UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Nonparametric Segmentation of Nonstationary Time Series

Laila Chahrazad Witzgall

Mestrado em Física
Especialização em Física Estatística e Não Linear

Dissertação orientada por:
Frank Raischel
José Pires Marques

2016

Declaration of Authorship

I, Laila Chahrazad Witzgall, declare that this thesis titled, 'Nonparametric segmentation of nonstationary time series' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: _____

Date: _____

"Look again at that dot. That's here. That's home. That's us. On it everyone you love, everyone you know, everyone you ever heard of, every human being who ever was, lived out their lives. The aggregate of our joy and suffering, thousands of confident religions, ideologies, and economic doctrines, every hunter and forager, every hero and coward, every creator and destroyer of civilization, every king and peasant, every young couple in love, every mother and father, hopeful child, inventor and explorer, every teacher of morals, every corrupt politician, every "superstar," every "supreme leader," every saint and sinner in the history of our species lived there-on a mote of dust suspended in a sunbeam."

Carl Sagan, Pale Blue Dot: A Vision of the Human Future in Space

Resumo

A análise de séries temporais trata do estudo de dados colectados durante determinado período de tempo. Uma série temporal consiste numa série de dados listados por ordem temporal, e é constituído por uma sequência de dados medida sucessivamente em intervalos de tempo equidistantes, ou não. O estudo de séries temporais é um campo vasto da estatística que se ramifica a várias áreas da ciência. A análise de séries temporais consiste em métodos de análise de dados com o objectivo de extrair elementos estatísticos e outras características relevantes e ocorre frequentemente no contexto da estatística, econometria, geofísica, meteorologia e outras áreas onde uma das principais motivações para o estudo destas séries temporais é a previsão. Uma grande parte dos sistemas complexos encontrados na vida real têm associados séries temporais empíricas que exibem graus variáveis de não-estacionariedade, como por exemplo medições da velocidade do vento, séries temporais financeiras, entre outros. Um processo estocástico estacionário tem como propriedade que a estrutura da média, variância e autocorrelação não se altera no tempo.

Um dos focos desta área de estudo é o tratamento de séries temporais não-estacionárias através de algoritmos de segmentação. A segmentação de séries temporais consiste em dividir a série em fragmentos, baseando a decisão de segmentação num critério pré-determinado no algoritmo. Neste trabalho explora-se um algoritmo de segmentação automática recursiva não-paramétrica baseado no teste estatístico de Kolmogorov-Smirnov para séries temporais não-estacionárias provenientes de processos complexos. A segmentação permite dividir a série temporal em fragmentos onde a estatística é idêntica, criando assim janelas de estacionariedade dentro de uma série não-estacionária. O teste de Kolmogorov-Smirnov é um teste totalmente não-paramétrico que avalia a igualdade de distribuições de probabilidade contínuas que pode ser utilizado para comparar uma amostra de dados com uma distribuição de probabilidade de referência, Teste de Kolmogorov-Smirnov para uma amostra, ou pode ser utilizado para comparar duas amostras de dados e neste caso designa-se por Teste de Kolmogorov-Smirnov para duas amostras. Este teste possibilita-nos testar se duas amostras pertencem a uma mesma distribuição sem necessidade de especificar qual, isto resulta da análise da diferença entre duas funções de distribuição cumulativas e observar em que ponto esta diferença absoluta é máxima. Esta diferença designa-se por distância de Kolmogorov-Smirnov.

Neste trabalho utiliza-se o conceito de teste de hipóteses que consiste numa categoria de inferência estatística fazendo parte de teoria da decisão. Um teste de hipóteses

inicia com a proposta de uma hipótese nula, em como um modelo probabilístico descreve as observações de determinada experiência. A questão abordada no teste tem como consequência dois possíveis resultados: aceitar ou rejeitar a hipótese nula. Neste caso estamos interessados em testar a existência de uma distribuição comum entre duas amostras de séries temporais. Dada a hipótese nula de que as duas amostras pertencem à mesma distribuição, podemos testar esta relativamente à hipótese alternativa de que as distribuições têm funções de distribuição cumulativas diferentes. Para cada amostra calcula-se a função de distribuição cumulativa e a diferença entre elas ponto a ponto. Comparamos esta distância e extraímos a distância máxima que constitui a estatística do teste, a distância de Kolmogorov-Smirnov entre as duas funções.

O algoritmo de segmentação para séries temporais aqui desenvolvido baseia-se nesta distância entre funções de distribuição cumulativas e funciona, em suma, da seguinte forma: dada uma série temporal e um ponteiro que se move sequencialmente em toda a série, a cada posição do ponteiro é feito um corte na amostra e são comparados os dois fragmentos resultantes. É calculada a estatística de Kolmogorov-Smirnov e quando o algoritmo percorre toda a série temporal é extraído o valor máximo desta estatística. Por sua vez, é nesta posição, onde o valor máximo é encontrado que o algoritmo propõe uma posição de corte da série temporal e compara este com a significância de uma possível posição de segmentação. Este processo é então aplicado iterativamente até não existirem mais propostas de posições de corte ou o fragmento testado tem tamanho inferior a um tamanho pré-determinado.

O objectivo principal do trabalho consistiu em caracterizar o algoritmo de segmentação testando séries temporais artificiais compostas por números aleatórios de distribuições diferentes, Gaussiana, log-normal e Cauchy. A escolha das distribuições de log-normal e de Cauchy foi motivada por estas serem classificadas como classes de distribuições com heavy tails, i.e., a cauda da distribuição é mais acentuada e decai como uma power-law. Muitas séries temporais de sistemas reais apresentam heavy tails e por esta razão é importante explorar o algoritmo e optimizá-lo para este tipo de distribuições. Explora-se também a função de probabilidade do teste de Kolmogorov-Smirnov e o critério de significância para amostras de tamanho muito grande. Este critério não se mostra adequado para o algoritmo aqui desenvolvido porque assume que as amostras comparadas pelo algoritmo são independentes o que não é o caso. O algoritmo tem como entrada uma série temporal que é dividida recursivamente em pares de fragmentos que são posteriormente comparados entre si o que torna os dados interdependentes e por este motivo utiliza-se um critério de significância adequado sugerido na literatura.

Numa fase seguinte realizam-se testes numéricos extensivos para avaliar a precisão e eficiência do algoritmo para diferentes distribuições, nomeadamente, Gaussiana, log-normal

e Cauchy. O algoritmo de segmentação de Kolmogorov-Smirnov mostra comportar-se bem mesmo quando testado em distribuições com heavy tails, caso em que o teste de Kolmogorov-Smirnov é, em teoria, menos sensível. Motivados por isto e procurando optimizar o desempenho do algoritmo para distribuições com *heavy-tails* introduzimos uma mudança ao algoritmo onde substituímos o teste de Kolmogorov-Smirnov pelo teste de Anderson-Darling que consiste em adicionar um termo com uma função de peso. Esta função de peso permite uma maior flexibilidade no sentido que mediante a escolha certa dá mais peso a determinada zona da distribuição, no nosso caso, a cauda. Com esta alteração ao algoritmo de segmentação analisou-se o comportamento do critério de significância que se mostrou continuar adequado. O algoritmo de segmentação com o teste de Anderson-Darling foi então aplicado a séries temporais construídas a partir de números aleatórios gerados a partir da distribuição de Cauchy e comparado à versão do algoritmo com o teste de Kolmogorov-Smirnov.

Em seguida analisa-se o desempenho do algoritmo de segmentação no espaço de parâmetros das distribuições para as duas versões do algoritmo, com o teste de Kolmogorov-Smirnov e com a introdução da modificação de Anderson-Darling. Com esta análise é possível fazer uma análise quantitativa do desempenho do algoritmo e deste modo estabelecer uma comparação entre ambas as vertentes do algoritmo. Esperava-se que a implementação do teste de Anderson-Darling otimizasse significativamente o desempenho do algoritmo quando aplicado a distribuições com *heavy-tails* verificando-se apenas uma ligeira melhoria quando aplicado a uma série temporal de Cauchy. Trabalho futuro poderia consistir em melhorar desempenho do algoritmo de segmentação em séries temporais com heavy tails, aumentando a sua sensibilidade nas caudas da distribuição.

Será interessante aplicar o algoritmo a medições empíricas de sistemas complexos reais tais como sistemas geofísicos ou sistemas socio-económicos situações onde distribuições com heavy tails têm um papel crucial. Será igualmente interessante analisar como é que o algoritmo de segmentação modificado, com a implementação do teste de Anderson-Darling ao invés do de Kolmogorov-Smirnov, aqui apresentado poderá auxiliar na distinção de diferentes regimes de parâmetros em séries temporais complexas de sistemas físicos reais, como por exemplo dados de mercados financeiros onde ocorrem tipicamente oscilações entre diferentes estados de mercado acompanhados de alterações nas distribuições de retorno, estruturas de correlação, expoentes de Hurst entre outros. Possivelmente em combinação com outras ferramentas estatísticas sensíveis a alterações nas quantidades previamente mencionadas, uma rotina de segmentação automatizada poderá ser útil, eficiente e uma assistência facilmente programável em *decision-making*.

Keywords: Time series, Kolmogorov-Smirnov Test, Anderson-Darling Test, *heavy-tails*

Abstract

Many empirical time series that arise in real-world complex systems are found to exhibit varying degrees of nonstationarity, such as atmosferic wind fields and financial time series. A nonparametric segmentation method for nonstationary time series has been implemented based on an existing algorithm using the statistical Kolmogorov-Smirnov test for equality of cumulative distribution functions. Starting from an automated segmentation algorithm based on the Kolmogorov-Smirnov distance for Gaussian, log-normal and Cauchy distributed random time series, we have attempted to characterize and improve the segmentation performance for heavy tailed time series. A time series can be understood to be composed of a series of reasonably long segments, for each of which its properties are stationary. The nonparametric segmentation algorithm presented here divides the time series recursively into segments and for each pair of resulting segments congruence of the respective empirical probability distribution function is asserted by the Kolmogorov-Smirnov test. The Kolmogorov-Smirnov test is weakly sensitive in the tails of the tested sample, when often these tail events are most interesting. For this reason we introduce a modification to the segmentation algorithm, replacing the Kolmogorov-Smirnov test with the Anderson-Darling test, incorporating a weight function to allow more flexibility in the test and account for the tails. In a primary phase we make a complete characterization of the segmentation algorithm and look to make improvements for heavy tailed distributions. We explore the Kolmogorov-Smirnov probability function for large sample sizes and the significance criterion for the classic Kolmogorov-Smirnov test and examine a proposed significance criterion suited for data that is not independent, which is our case because we start from an integral time series that is recursively divided into fragments and compared. In a final phase we investigate the efficiency and performance range of the segmentation algorithm with the Kolmogorov-Smirnov test for Gaussian, log-normal and Cauchy distributed time series. We implement the Anderson-Darling test and establish a comparison with the Kolmogorov-Smirnov based segmentation algorithm for heavy tailed distributed time series.

Keywords: Time series, Kolmogorov-Smirnov Test, Anderson-Darling Test, heavy-tails

Acknowledgements

I would like to thank my thesis supervisor, Frank Raischel, who was always available whenever I needed help in my research and writing. I will be eternally thankful for his supervision and friendship. I also want to thank my second thesis supervisor, José Pires Marques, for his patience, comprehension, availability and for making this possible, it was a pleasure to work with him throughout my Bachelor and Masters degrees. Also, I am thankful to Sílvio M. Duarte Queirós, one of the authors of the paper this thesis was based on, for clarifying some questions we had about notation. I also want to thank Diogo Sousa for the help in coding the algorithm in its initial phase.

I want to show my gratitude to the Câmara Municipal de Aljezur, namely Maria de Fátima Neto and José Amarelinho for granting me a scholarship helping me financially.

I thank my friends from Lisbon, in particular Sara Freire, Rita Freire, Catarina Ramos and Rosa Santos for supporting me, believing in me and giving me strength.

Also I thank my friends in Aljezur, my beautiful hometown, specially João Serafim for being my friend for 21 years, Celine Gisin who recently stepped into my life and made it brighter with her sunkissed friendship, gave me strength in the final steps and made me believe the world is a good place, also I want to thank Rute Fernandes for being like a second mother and always encouraging me.

Finally, I am profoundly grateful to my mother, Saïda Hanine, for always supporting me and helping me through everything. Thanks to her continuous encouragement throughout all my studies this thesis was possible. I also express greatest gratitude to my dear aunt, Rajae Berrada for motivating me and helping me as if I was her own daughter.

Laila Chahrazad Witzgall

Contents

List of Figures

Abbreviations

AD	**A**nderson-**D**arling
CDF	**C**umulative, **D**istribution **F**unction
iid	**i**ndependent, **i**dentically **d**istributed
iif	**i**f and only **if**
KS	**K**olmogorov-**S**mirnov
PMF	**P**robability **M**ass **F**unction
PDF	**P**robability **D**istribution **F**unction

Symbols

$P[.]$	Probability measure
S	Sample space
s	Sample outcome
X	Random Variable
$X(s)$	function mapping s to its corresponding random variable
$\{X = x\}$	Set of sample points $s \in S$ for which $X(s) = x$
$P_X(x)$	Probability mass function
$F_X(x)$	Cumulative distribution function
$f_X(x)$	Probability distribution function
$\mathbb{E}[X]$	Expected value of X
$\mathrm{Var}[X]$	Variance of X
σ	Standard deviation
σ_X	Standard deviation for the log-normal distribution
σ_Y	Scale parameter for the log-normal distribution
μ_Y	Location parameter for the log-normal distribution
γ	Scale parameter of the Cauchy distribution
x_0	Location parameter of the Cauchy distribution
X_n	Random sequence
t	Time
τ	Small time interval
$X(t)$	Stochastic process
$W(t)$	Brownian motion process
$C_X(t, \tau)$	Autocovariance of $X(t)$
$R_X(t, \tau)$	Autocorrelation of $X(t)$
H_0	Null hypothesis

H_1	Alternative hypothesis
α	Significance level
$P_0 = 1 - \alpha$	Confidence level
N_e	Effective number of data points $\left(\frac{N_1+N_2}{N_1 \cdot N_2}\right)^{-1}$
N_{x_i}	Number of data points in sample i
D_{AD}	Anderson-Darling Distance
$\psi(t)$	Weight function
t	Empirical CDF for the AD test
i	Position in the data array
i_p	Pointer position
$i_{\max}$	Position that maximizes D_{KS}
D_{KS}	Kolmogorov-Smirnov distance
$D_{KS}(i)$	KS statistic at position i
$D(i)$	Length weighted Kolmogorov-Smirnov distance, $D = D_{KS} \cdot \sqrt{N_e}$
$D_{\max}$	Maximal distance, $D(i_{\max})$
$D_{KS}(i_{\max})$	KS distance at $i_{\max}$
$D_{\max}^{\text{crit}}$	Cut acceptance criterion for the segmentation algorithm
Q_{KS}	KS probability function
λ	Weighted KS distance, D_{KS}
λ^c	Critical λ values
R	Number of realisations
m	Segment sizes of the time series in the performance tests for the KS segmentation algorithm
$\delta_{(1/e)}$	$1/e$ width of the PDF

Dedicated to my mother.

Chapter 1

Introduction

1.1 Motivation

Many empirical time series from complex systems are found to exhibit varying degrees of non-stationarity. Examples are measurements from atmospheric wind fields [1], atmospheric pollutant concentrations and financial time series [2]. Although the definition of non-stationarity is not unique, it is generally understood to refer to significant differences between probability distribution functions (PDF) measured over partial segments of a time series. The time series can then be understood to be composed of a series of reasonably long segments, for each of which its properties are stationary. Recently, a nonparametric segmentation algorithm has been presented that is capable of dealing with time series from atmospheric wind measurements and other complex systems [3]. For this purpose, the time series is recursively divided into pairs of slices. For each pair, congruence of the respective empirical PDF's is asserted by the Kolmogorov-Smirnov Test [4]. It has been shown [3] that the results are superior to a widely-used statistical test based on the Student's t-test [5]. The algorithm is based on the KS test which shows accuracy in detecting differences between CDF's when dealing with distributions that do not have relevant information in the tails. This motivates us to look for an improvement that would work well for heavy tailed distributions where we apply a modification to the KS test, the Anderson-Darling test.

1.2 Probability and Stochastic Processes

This introductory chapter covers relevant theory of probability and stochastic processes based on [6] and [7]. We give an introduction to probability of discrete and continuous random variables in Subsections 1.2.1 and 1.2.2, respectively, to familiarise the reader

with concepts and notation. In Subsection 1.2.1 we define a discrete random variable and introduce its probability mass function (PMF) and cumulative distribution function (CDF) along with mathematical definitions for average, variance and standard deviation. Next, in Subsection 1.2.2 we move to continuous random variables introducing the concept of the probability density function and expected values and finally give an overview of relevant classes of distributions to this work, namely, the Gaussian distribution in 1.2.2.3, log-normal distribution in 1.2.2.4 and Cauchy distribution in 1.2.2.5. Afterwards, in Section 1.4 we turn our attention to stochastic processes and stationarity. Section 1.5 contains an introduction to two categories of statistical inference offering an overview on hypothesis testing, parametric and nonparametric tests leading to our main topic, the Kolmogorov-Smirnov Test in 1.5.4 and the Anderson-Darling Test in 1.6.1. Finally we describe in detail the segmentation algorithm to be used throughout the present work in Section 1.7

1.2.1 Discrete Random Variables

In probability theory exist two types of random variables, discrete random variables and continuous random variables which differ in the set of possible values they can take, that is, a discrete random variable can take only a countable number of possible values while a continuous random variable can take any value in a given interval. We start with defining discrete random variables and related important concepts such as the cumulative distribution function, averages, variance and standard deviation. A discrete random variable can take only a countable number of values. More precisely, X is *discrete* if there exists a finite or countable set $S \subset \mathbb{R}$ such that $P[X \in S] = 1$, i.e, if we know that the only values X can take are those existing in S.

1.2.1.1 Random Variable and Probability Density

We define *random variable* as follows

Definition 1.2.1. A **random variable** consists of an experiment with a probability measure $P[\cdot]$ defined on a sample space S and a function that assigns a real number to each outcome in the sample space of the experiment [7].

We characterise the random variable X by the function $X(s)$ that maps the sample outcome s to the matching value of the random variable. We write that $\{X = x\}$ to say that there is a set of sample points $s \in S$ for which $X(s) = x$. That is, [7]

$$\{X = x\} = \{s \in S \mid X(s) = x\}. \tag{1.1}$$

The possible values for *discrete random variable* form a countable set and the experiments shape discrete sample spaces. On the other hand if the random variable takes any real number it is denoted as a *continuous random variable* and the underlying experiment takes a continuous sample space. Continuous random variables will be presented in Subsection 1.2.2.

We define a *discrete random variable* as

Definition 1.2.2. X is a **discrete** random variable if the range of X is a countable set $S_X = \{x_1, x_2, \dots, \}$ [7].

For a discrete random variable the probability measure that gives us probabilities of the possible values for a random variable is the probability mass function (PMF), $P_X(x)$. The PMF contains all information about the underlying probability model and is defined as follows

Definition 1.2.3. The **probability mass function** (PMF) of the discrete random variable X is given by [7]

$$\mathrm{P}_X(x) = \mathrm{P}[X = x]$$

The PMF, $P_X(x)$, is a real valued function and gives the probability of the event $X = x$. Note that $X = x$ is an event consisting of all possible outcomes s of the underlying experiment for which $X(s) = x$. The following Theorem states three basic important properties concerning the PMF $P_X(x)$ of a discrete random variable X,

Theorem 1.2.1. For a discrete random variable X with PMF $P_X(x)$ and range S_X [7]:

(i) For any x, $P_X(x) \geq 0$.

(ii) $\sum_{x \in S_X} P_X(x) = 1$.

(iii) For any event $B \subset S_X$, the probability that X is in the set is

$$P[B] = \sum_{x \in B} P_X(x).$$

In the next Section we introduce the cumulative distribution function of a random variable which is closely related to the probability mass function.

1.2.1.2 Cumulative Distribution Function

The cumulative distribution function, CDF, such as the PMF, of a discrete random variable contains complete information about the probability model of the random variable.

The two functions are closely related and can be obtained from each other. The CDF gives us the probability that a random variable takes value less than or equal to x, as stated by Definition 1.2.4:

Definition 1.2.4. The **Cumulative Distribution Function** (CDF) of a random variable X is [7]

$$F_X(x) = P[X \leq x].$$

All random variables have a CDF but the probability mass function, PMF, is only defined for discrete random variables. Theorem 1.2.2 states important properties of the CDF:

Theorem 1.2.2. For any discrete random variable X with range S_X satisfying $x_1 \leq x_2 \leq \ldots$, [7]

(i) $F_X(-\infty) = 0$ and $F_X(\infty) = 1$, i.e, from left to right on the x-axis, $F_X(x)$ starts at zero and ends at one.

(ii) For all $x' \geq x$, $F_X(x') \geq F_X(x)$, i.e, the CDF never decreases and goes from left to right.

(iii) For $x_i \in S_X$ and ϵ and arbitrarily small positive number,

$$F_X(x_i) - F_X(x_i - \epsilon) = P_X(x_i).$$

This means that there is a jump (discontinuity) at each value of $x_i \in S_X$. The height of this jump at x_i is $P_X(x_i)$.

(iv) Between jumps, the graph of the CDF is a horizontal line

$$F_X(x) = F_X(x_i) \, \forall \, x : \, x_i \leq x < x_{i+1}.$$

From the definition of the CDF, given by Definition 1.2.4, follows an important theorem that the difference between the CDF evaluated at any two points a and b is the probability that the random variable X takes a value between these two points:

Theorem 1.2.3. For all $b \geq a$, [7]

$$F_X(b) - F_X(a) = P[a < x \leq b].$$

Next we introduce the concept of averages and the inherent definitions.

1.2.1.3 Averages

The average value, also known as expected value or simply expectation of a random variable is denoted by $\mathbb{E}[X]$. It is a parameter that plays an important role in characterising a probability distribution. The familiar term of arithmetic mean is simply given by the sum of a certain number of measurements divided by the number of terms in the sum, but in statistics we add weight terms $P_X(x)$ thus the average value can be interpreted as a weighted average

Definition 1.2.5. The **expected value** of X is [7]

$$\mathbb{E}[X] = \mu_X = \sum_{x \in S_X} x\, P_X(x).$$

In the next subsection we introduce two important measures of dispersion of a distribution, the variance and standard deviation.

1.2.1.4 Variance and Standard Deviation

The most important measures of dispersion are the standard deviation and the variance. Dispersion describes how spread out a data set is and standard deviation is the most common measure, it tells us how far apart are the numbers from the mean value. The variance of a random variable X describes the squared difference between X and its expected value.

Definition 1.2.6. The **variance** of a random variable X is [7]

$$\mathrm{Var}[X] = \mathbb{E}\left[(X - \mu_X)^2\right]$$

The standard deviation can be positive and negative so it is needed to square the values to ensure the values do not cancel each other after adding them up, hence we say that the standard deviation σ is given by the square root of the variance as stated in the following definition

Definition 1.2.7. The **standard deviation** of a random variable X is [7]

$$\sigma_X = \sqrt{\mathrm{Var}[X]}.$$

Because $(X - \mu_X)^2$ is a function of X, $\mathrm{Var}[X]$ can be computed as follows

$$\mathrm{Var}[X] = \sigma_X^2 = \sum_{x \in S_X} (x - \mu_X)^2 \, P_X(x).$$

The above expression for the variance, $\mathrm{Var}[X]$, can be expanded and we are led to the following theorem[7]

Theorem 1.2.4.

$$\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu_X^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \,.$$

The quantities $\mathbb{E}[X]$ and $\mathbb{E}[X^2]$ are denoted *moments* of the random variable X. Moments of a distribution are written in terms of the expectation value of the random variable X, as follows [7]:

Definition 1.2.8. For a random variable X:

(i) The nth **moment** is $\mathbb{E}[X^n]$.

(ii) The nth **central moment** is $\mathbb{E}\left[(X - \mu_X)^n\right]$

Hence, $\mathbb{E}[X]$ is the *first moment* of X, $\mathbb{E}[X^2]$ is the *second moment* of X and $\mathrm{Var}[X]$ is a central moment of X. Theorem 1.2.4 shows us that the variance of X, $\mathrm{Var}[X]$, is given by the difference between the second moment of X and the square of the first moment, or the nth central moment.

In Section 1.2.1 we introduced the notion of discrete random variables and their cumulative distribution function (CDF), average, variance and standard deviation measures. In the next section we turn our attention to continuous random variables and related important concepts such as the probability density function, expected values and introduce distributions that are relevant in Chapter 2.

1.2.2 Continuous Random Variables

In the previous section we introduced discrete random variables but many real world experiments are described by continuous random variables. In this section we analyse random variables that range over a continuous interval of numbers containing all real numbers within this interval. We say that a random variable is continuous if its range is a continuous interval or, equivalently, if its distribution function is continuous everywhere. In the case of discrete random variables we used the PMF to make a complete description of the underlying probability model but in the case of continuous random variables it is not possible to define a PMF but we make use of another very useful probability model suitable for this kind of random variables, the CDF.

The **Cumulative Distribution Function** (CDF) of a random variable X is defined in Definition 1.2.4 and the most important properties of the CDF are listed in Theorem 1.2.2 and apply to all types of random variables [7].

Definition 1.2.9. X is a **continuous random variable** if the CDF $F_X(x)$ is a continuous function.

Next we approach the concept of probability density function (PDF) and its properties for a continuous random variable.

1.2.2.1 Probability Density Function

The *probability density* is defined as the first derivative of the CDF, which should be differentiable everywhere, and contains all relevant information about a continuous random variable. The probability density function (PDF) is defined as the first derivative of $F_X(x)$ of X, when it exists, and is denoted by f_X [7]:

Definition 1.2.10. The **probability density function** (PDF) of a continuous random variable X is

$$f_X(x) = \frac{d\,F_X(x)}{dx}$$

The PDF contains all information about the continuous random variable at study. Important properties of the PDF follow from Definition 1.2.10 and the properties of CDFs and are listed in Theorem 1.2.5 [7]:

Theorem 1.2.5. For a continuous random variable X with PDF $f_X(x)$,

(i) $f_X(x) \geq 0,\ \forall\, x.$

(ii) $F_X(x) = \int_{-\infty}^{x} f_X(u)\, du.$

(iii) $\int_{-\infty}^{\infty} f_X(x)\, dx = 1.$

The first property of Theorem 1.2.5 is a consequence of the distribution function being a nondecreasing function hence, the PDF is always nonnegative. Next we turn to the notion of expected values and moments of continuous random variables.

1.2.2.2 Expected Values

The expected value, also known as expectation or first moment, of a random variable is the probability weighted average of all possible values. The concepts we defined for discrete random variables in Section 1.2.1 carry over straightforwardly to continuous

random variables. For continuous random variables the expectation value is defined as the integral of the random variable with respect to its probability measure. Let $\mathbb{E}[X]$ denote the expected value of a continuous random variable X with PDF $f_X(x)$ [7]:

Definition 1.2.11. The **expected value** of a continuous random variable X is

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x\, f_X(x)\, dx.$$

Many properties of expected values of discrete random variables apply to continuous random variables and we can summarise them in terms of expected values in the following Theorem [7]:

Theorem 1.2.6. For any random variable X,

(i) $\mathbb{E}[X - \mu_X] = 0$,

(ii) $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$,

(iii) $\mathrm{Var}[X] = \mathbb{E}[X^2] - \mu_X^2$

(iv) $\mathrm{Var}[aX + b] = a^2\mathrm{Var}[X]$

The second moment of X and the variance of X, $\mathrm{Var}[X]$, are given by [7]

$$\mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} x^2\, f_X(x)\, dx, \qquad \mathrm{Var}[X] = \int_{-\infty}^{\infty} (x - \mu_X)^2\, f_X(x)\, dx. \tag{1.2}$$

We have introduced concepts of probability theory and in the next sections we introduce a number of distributions and their key properties that are relevant in the work done in Chapter 2.

1.2.2.3 Gaussian Distribution

Many random variables in physical contexts are distributed in such a way to present a normal or Gaussian distribution. A random variable X with PDF given by Definition 1.2.12 is defined as a normal or Gaussian random variable with mean μ and variance σ^2 or standard deviation σ. An example of the PDF of normally distributed random variable with different means μ and standard deviations σ are shown in Fig. 1.1

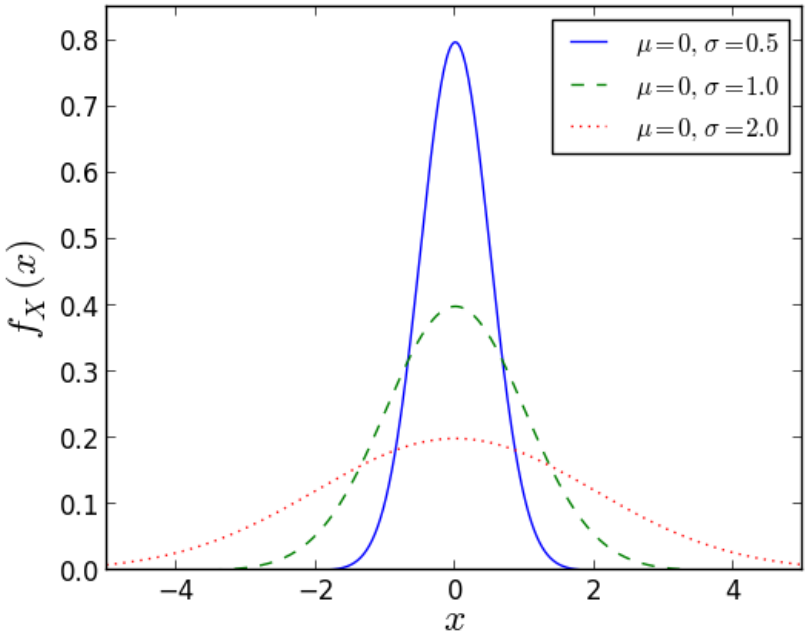


FIGURE 1.1: Probability density functions for Gaussian distribution with expected value μ and standard deviation σ.

As we see from Fig. 1.1 the Gaussian probability distribution has a bell shape and is symmetrical about the mean μ.

Definition 1.2.12. X is a Gaussian (μ, σ) random variable if the PDF of X is

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{\frac{-(x-\mu)^2}{2\sigma^2}},$$

where the parameter μ can be any real number and the parameter $\sigma > 0$.

The graph of $f_X(x)$ has a bell shape where the center of the bell is located at $x = \mu$ while the standard deviation σ reflects the width of the bell translating the dispersion of the data. For small σ the bell is narrow with a high peak. If σ is large the bell is wider with a low flat peak as one can see in Fig. 1.1. The height of the peak is given by $1/\sigma\sqrt{2\pi}$.

Theorem 1.2.7. If X is a Gaussian (μ, σ) random variable,

$$\mathbb{E}[X] = \mu, \qquad \mathrm{Var}[X] = \sigma^2.$$

Theorem 1.2.8. If X is a Gaussian (μ, σ) random variable, $Y = aX + b$ is Gaussian $(a\mu + b, a\sigma)$

The above theorem states that any linear transformation of a Gaussian random variable results in another Gaussian random variable. With this Theorem we can relate the

properties of an arbitrary Gaussian random variable to the properties of a specific random variable.

Definition 1.2.13. The **standard normal random variable** Z is the Gaussian $(0, 1)$ random variable.

Theorem 1.2.9. The CDF of the standard normal random variable Z is

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\, du.$$

Theorem 1.2.10. If X is a Gaussian (μ, σ) random variable, the CDF of X is

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right).$$

The probability that X is in the interval $(a, b]$ is

$$P[a < X \leq b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

1.2.2.4 Log-normal Distribution

The log-normal distribution has a heavy-tailed form and is a standard distribution in finances to model stock price movements and fluctuations. We present the mathematical definition of the log-normal distribution and key concepts.

A positive random variable X is said to have log-normal distribution if its logarithm [6]

$$Y = \ln X$$

has the normal distribution, i.e.:

$$f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_Y} e^{\left[-\frac{(y-\mu_Y)^2}{2\sigma_Y^2}\right]}, \qquad -\infty < y < \infty. \tag{1.3}$$

In order to find the PDF of the log-normal random variable X we use that $dy = dx/x$ and $f_Y(y)dy = f_X(x)dx$:

Definition 1.2.14. X is a log-normal random variable for which $Y = \ln X$ has a normal distribution with location parameter μ_Y and scale parameter σ_Y. The PDF is given by

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma_Y x} e^{\left[-\frac{(\ln x-\mu_Y)^2}{2\sigma_Y^2}\right]},$$

where $0 \leq x < \infty$.

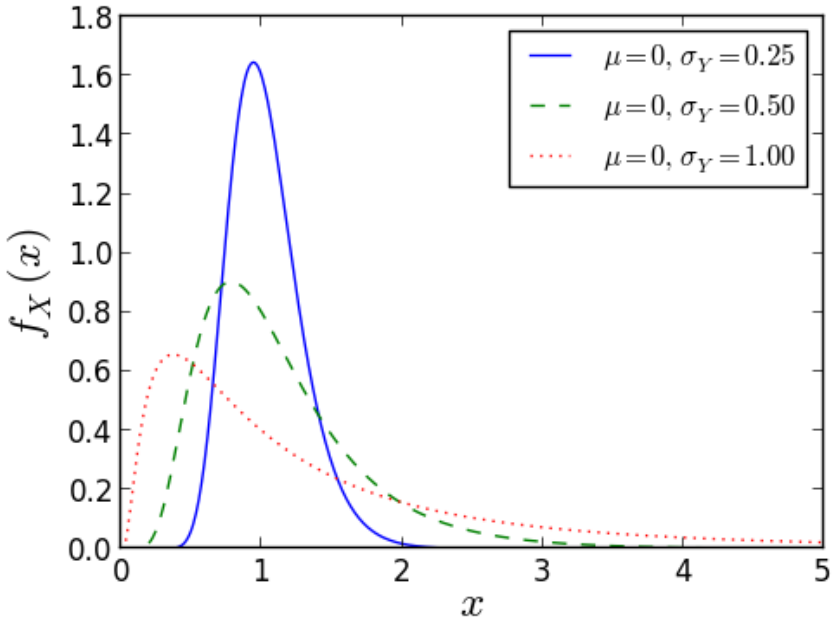


FIGURE 1.2: Probability density functions for log-normal distribution with location parameter $\mu = 0$ and scale parameters $\sigma_Y = 0.25, 0.5\,1.0$.

To obtain the expected value of X we make use of the moment-generating function of the normal random variable Y given by, [6]:

$$\mathbb{E}[e^{tY}] = e^{\left(\mu_Y t + \frac{\sigma_Y^2 t^2}{2}\right)}. \tag{1.4}$$

Then

Definition 1.2.15. The expected value of the log-normal random variable X is given by the first moment

$$\mu_X = \mathbb{E}[X] = \mathbb{E}[e^Y] = e^{\left(\mu_Y + \frac{\sigma_Y^2}{2}\right)} \tag{1.5}$$

In the same way, the second moment of X is

$$\mathbb{E}\left[X^2\right] = \mathbb{E}\left[e^{2Y}\right] = e^{\left(2\mu_Y + 2\sigma_Y^2\right)} = \mu_X^2 e^{\sigma_Y^2} \tag{1.6}$$

Therefore, the variance of the log-normal variable is given by

$$\begin{aligned}\sigma_X^2 &= e^{\left(2\mu_Y+2\sigma_Y^2\right)} - \left[e^{\left(\mu_Y+\frac{\sigma_Y^2}{2}\right)}\right]^2 \\ &= e^{\left(2\mu_Y+\sigma_Y^2\right)}\left[e^{\left(\sigma_Y^2\right)}-1\right] \\ &= \mu_X^2\left[e^{\left(\sigma_Y^2\right)}-1\right] \end{aligned} \tag{1.7}$$

From Definition 1.2.15 and 1.7 we obtain expressions for the mean and the variance of Y in terms of those of X:

$$\mu_Y = \ln \mu_X - \frac{1}{2}\ln\left(1+\frac{\sigma_X^2}{\mu_X^2}\right), \tag{1.8}$$

and

$$\sigma_Y^2 = \ln\left(1+\frac{\sigma_X^2}{\mu_X^2}\right). \tag{1.9}$$

1.2.2.5 Cauchy Distribution

The Cauchy distribution is unimodal and symmetric with heavy tails. The PDF is symmetric about the location parameter x_0 [8].

Definition 1.2.16. The probability distribution for Cauchy random variable is

$$f_X(x) = \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}, \qquad -\infty < x < \infty \tag{1.10}$$

where x_0 is the location parameter, specifying the location of the peak of the distribution, and γ is the scale parameter which specifies the half-width at half-maximum.

The Cauchy distribution has no mean, variance or higher moments defined. Its mode and median are well defined and both equal to the location parameter x_0.

Definition 1.2.17. The special case when $x_0 = 0$ and $\gamma = 1$ is called the standard Cauchy distribution with PDF

$$f_X(x) = \frac{1}{\pi\left(1+x^2\right)},$$

The class of Cauchy distributions is closed under linear transformations.

Theorem 1.2.11. If X is a Cauchy distribution (x_0, σ) random variable, $Y = aX + b$ also is Cauchy $(|a|\, x_0, a\sigma + b)$.

The PDFs of the Cauchy random variable for selected values of the scale parameter γ and location parameter x_0 are shown in Fig. 1.3.

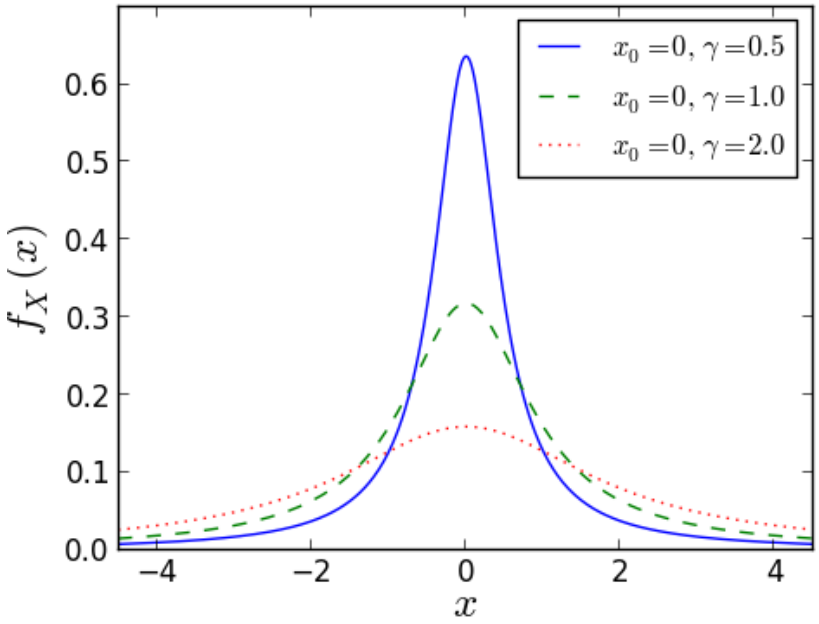


FIGURE 1.3: Probability density functions for Cauchy distribution with different location and scale parameters.

In Fig. 1.3 comparing the curves with scale parameters $\gamma = 0.5$ and $\gamma = 2$ we can observe that the higher the value of the scale parameter the heavier are the tails.

1.3 Power Laws

When the probability of measuring a particular value of some quantity varies inversely as a power of that same value, this quantity is said to follow a power law, also commonly known in literature as *Zipf's law* or the *Pareto distribution*. Power laws appear widely in physics, biology, earth and planetary sciences, economics and finance, computer science, demography and social sciences [9].

Let $p(x)\,\mathrm{d}x$ be the fraction of some sample between x and $x + \mathrm{d}x$. If the histogram is a straight line on log-log scales, then $\ln p(x) = -\alpha \ln x + c$, where α and c are constants. Taking the exponential of both sides we write

$$p(x) = Cx^{-\alpha}, \quad x > x_m \tag{1.11}$$

with $C = e^c$, α is called the *exponent* of the power law and x_m the minimum of the distribution. Distributions with the form of Eq. (1.11) are said to follow a *power-law*.

1.3.1 Measuring Power Laws

The standard strategy of detecting a power law is that a histogram of a quantity with a power law distribution appears as a straight line when plotted on logarithmic scales. But this is, in most cases, a poor way to proceed [9].

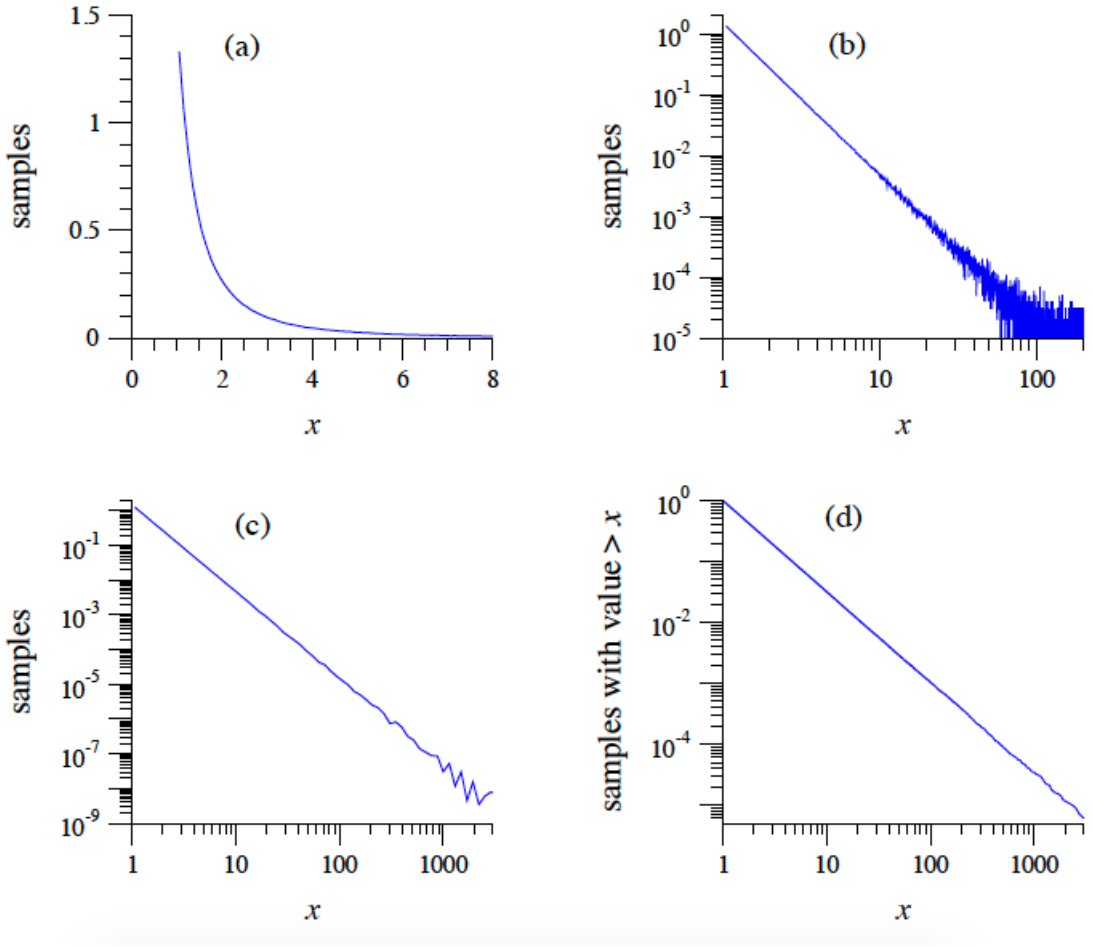


FIGURE 1.4: Power laws and logarithmic scales. (a) Histogram of the set of 1 million random numbers described in the text, which have a power-law distribution with exponent $\alpha = 2.5$. (b) The same histogram on logarithmic scales. Notice how noisy the results get in the tail towards the right-hand side of the panel. This happens because the number of samples in the bins becomes small and statistical fluctuations are therefore large as a fraction of sample number. (c) A histogram constructed using "logarithmic binning". (d) A cumulative histogram or rank/frequency plot of the same data. The cumulative distribution also follows a power law, but with an exponent of $\alpha - 1 = 1.5$. From [9].

Considering Fig. 1.4 we see an artificial data set composed of one million random real numbers generated from a power law probability distribution given by Eq. (1.11) with

exponent chosen to be $\alpha = 2.5$. Fig. 1.4 (a) shows a normal histogram of the numbers binned in equal bins of size 0.1. This produces a smooth curve on linear scale. Next, to reveal the power law form of the distribution we plot the histogram on a logarithmic scale, and when we do this we see the straight line form of the power law distribution displayed in (b). However, this plot is not a very good one because on the right-hand end the distribution is quite noisy due to sampling errors. We can not simply cut out the data in the tail of the curve because there os often useful information in those data, in fact, many distributions follow a power law only in the tail and a good solution is logarithmic binning. This means that bin sizes grow exponentially and thus the tail of the distribution gets more samples than it would if bin sizes were fixed and this results in reducing statistical errors in the tail region. Finally in (c) we can see that the straight line power law form of the histogram is much clearer [9].

Even with logarithmic binning there is still some noise in the tail. Suppose the bottom of the lowest bin is at $x_{\min}$ and the ratio of the widths of successive bins is a. Then the kth bin extends from $x_{k-1} = x_{\min}a^{k-1}$ to $x_k = x_{\min}a^k$ and the expected number of samples falling in this interval is [9]:

$$\begin{aligned}\int_{x_{k-1}}^{x^k} p(x)\,\mathrm{d}x &= C\int_{x_{k-1}}^{x^k} x^{-\alpha}\,\mathrm{d}x \\ &= C\frac{a^{\alpha-1}-1}{\alpha-1}\left(x_{\min}a^k\right)^{-\alpha+1}.\end{aligned}$$

As long as $\alpha > 1$, the number of samples per bin goes down as k increases and the bins in the tails will have more statistical noise than those that precede them. Most power law distributions that occur in nature have $2 \leq \alpha \leq 3$, so noisy tails are the norm.
Another way, and in many ways superior, of plotting the data is to calculate the cumulative distribution function (CDF), see Fig. 1.4 (d). In this way we do not throw away any information. CDFs which follow a power law are sometimes said to follow *Zipf's law* or a *Pareto distribution*. In practical situations we are interested to estimate the exponent α from observed data, which can be done by employing the formula

$$\alpha = 1 + n\left[\sum_{i=1}^{n}\ln\frac{x_i}{x_{\min}}\right]^{-1}. \tag{1.12}$$

Here the quantities x_i, $i = 1, \ldots, n$ are the measured values of x and $x_{\min}$ is the minimum value of x. In fact, in practical situations $x_{\min}$ usually corresponds not to the smallest value of x measured but to the smallest for which the power law behaviour holds [9].

1.4 Stochastic Processes

In this section we introduce the concept of stochastic processes, we give a brief introduction and an overview of stationarity of a stochastic process, we introduce basic concepts of particular importance such as the *autocorrelation function* and the *autocovariance function* of a stochastic process. These functions are useful description of the time structure of a process, just as the expected value and variance are useful characterisations of the amplitude structure of a random variable [7]. In stochastic processes each observation corresponds to a time dependent function. In many real-world experiments of probability involve taking observations from some physical system over some time interval and typically the time series under investigation in the present work can be modelled as stochastic processes.

1.4.1 Definitions

We start with the definition of a stochastic process, also denoted by random process, and related key concepts [7].

Definition 1.4.1. A **stochastic process** $X(t)$ consists of an experiment with a probability measure $P[\cdot]$ defined on a sample space S and a function that assigns a time function to each outcome s in the sample space of the experiment.

Thus, a stochastic process maps a random function of time to each outcome s of some random experiment [7].

Definition 1.4.2. A **sample function** $x(t, s)$ is the time function associated with the experimental outcome s.

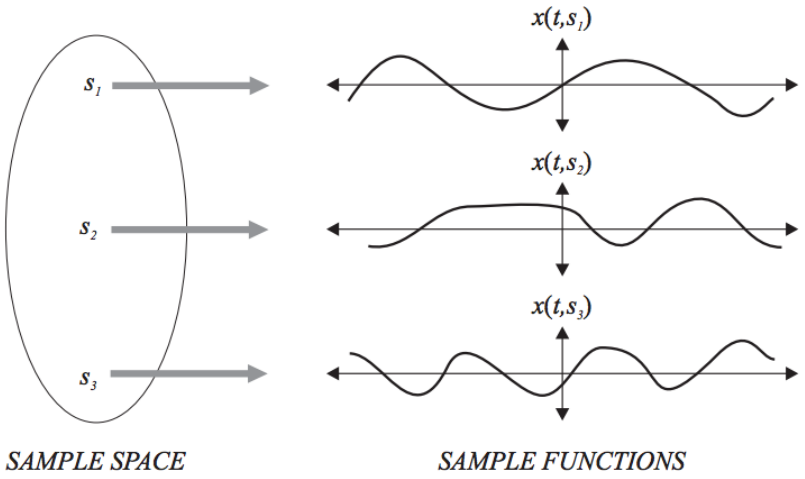


FIGURE 1.5: Conceptual representation of a random process. A stochastic process maps a random function of time, a sample function $x(t, s)$, to each outcome s of some random experiment. From [7].

A sample function is a function of time which is associated with the outcome of an experiment governed by a stochastic process. In Fig 1.5 we can observe the relation between the sample space of an experiment and the collection of sample functions of the underlying random process. The collection of sample functions is known as the *ensemble* of a stochastic process and we define it in the following way [7]:

Definition 1.4.3. The **ensemble** of a stochastic process is the set of all possible time functions that can result from an experiment.

Next we explore types of stochastic processes that may arise in complex systems, such as *Discrete-Value and Continuous-Value Processes* and *Discrete-Time and Continuous-Time Processes.*

1.4.2 Types of Stochastic Processes

In the previous sections for random variables we saw that they can be classified into two categories, discrete and continuous. In the same way we can define different categories for stochastic processes. These categories depend on the range of values taken at a time instant and the time instants themselves at which changes in the stochastic process occur [7].

The values taken by $X(t)$ are called *states* and the set of all possible values encompass the *state space* of the random process. If the set of all possible values of $X(t)$, the state space, is continuous then the process is classified as a *continuous-value random process* on the

other hand, if the state space of the stochastic process is discrete then it is a **discrete-value process**. If the time index T takes discrete values the process is a *discrete-time random process*, also called *random sequence* and denoted by $\{X_n, n = 1, 2, \ldots\}$, while if T is continuous we have a *continuous-time random process* [7] [10].

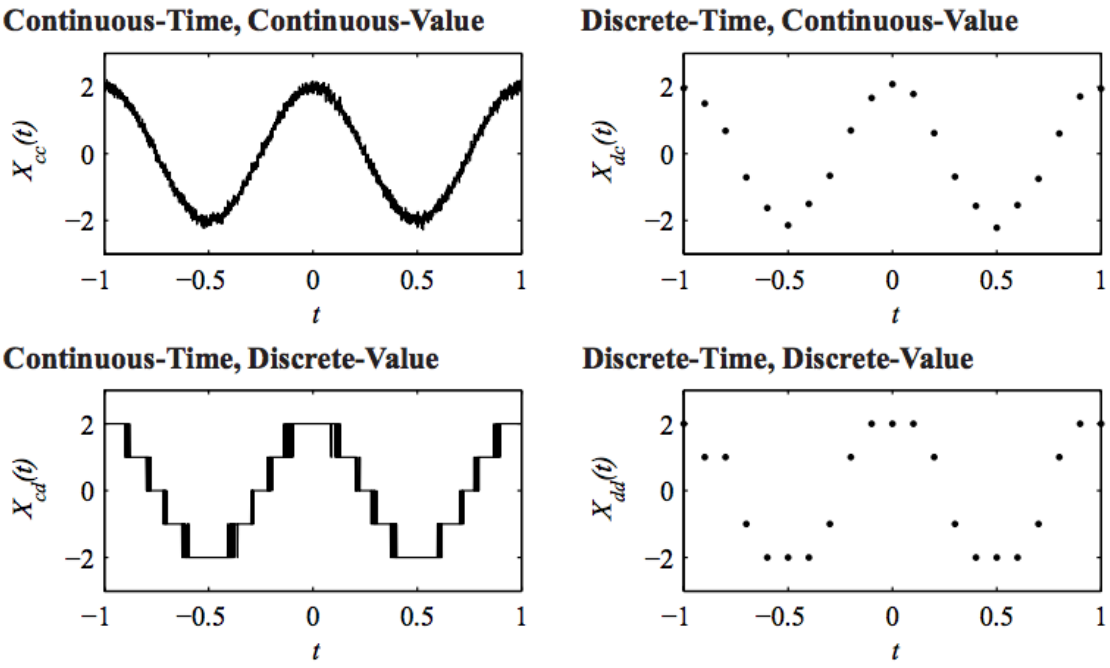


FIGURE 1.6: Sample functions of four kinds of stochastic processes. $X_{cc}(t)$ is a continuous-time, continuous-value process. $X_{dc}(t)$ is discrete-time, continuous-value process obtained by sampling $X_{cc}(t)$ every 0.1 seconds. Rounding $X_{cc}(t)$ to the nearest integer yields $X_{cd}(t)$, a continuous-time, discrete-value process. Lastly, Xdd (t), a discrete-time, discrete-value process, can be obtained either by sampling $X_{cd}(t)$ or by rounding $X_{dc}(t)$. From [7].

In Fig. 1.6 we see that combinations of continuous/discrete time and continuous/discrete value result in four categories [7].

Definition 1.4.4 (Discrete-Value and Continuous-Value Processes). $X(t)$ is a **discrete-value process** if the set of all possible values of $X(t)$ at all times t is a countable set S_X; otherwise $X(t)$ is a **continuous-value process**.

Definition 1.4.5 (Discrete-Time and Continuous-Time Processes). The stochastic process $X(t)$ is a **discrete-time process** if $X(t)$ is defined only for a set of time instants, $t_n = nT$, where T is a constant and n is an integer; otherwise $X(t)$ is a **continuous-time process**.

Definition 1.4.6. A **random sequence** X_n is an ordered sequence of random variables $X_0, X_1, \ldots$.

An example of a random sequence is the situation of tossing several times, consecutively, a fair coin.

1.4.3 Random Variables from Random Processes

Suppose we observe a stochastic process at a some time instant t_1. Each time we perform the experiment, we obtain a sample function $x(t, s)$ and that function specifies the value of $x(t_1, s)$. Each time we perform the experiment, we have a new s and we observe a new $x(t_1, s)$. Therefore, each $x(t_1, s)$ is a sample value of a random variable. We will use the notation $X(t_1)$ for this random variable [7].

With respect to a single random variable X we saw that all properties of a random variable X are determined from its PDF $f_X(x)$. In the same manner, for a pair of random variables X_1, X_2, we need the concept of joint PDF, $f_{X_1,X_2}(x_1, x_2)$. For a random processe, if we sample a process $X(t)$ at k time instants $t_1, \ldots, t_k$, we obtain the k-dimensional random column vector $\mathbf{X} = [X(t_1) \ldots X(t_k)]'$, where prime denotes the transpose matrix [7].

For a random variable X, we describe X by means its PDF $f_X(x)$ without specifying the underlying experiment. In the same way, knowledge of the joint PDF $f_{X(t_1),\ldots,X(t_k)}(x_1, \ldots, x_k)$ will allow us to describe a random process without specifying the underlying experiment.

1.4.4 Independent, Identically Distributed Random Sequences

Given the fact that a random process is viewed as an ensemble of random variables indexed by time we can broaden the concept of independent random variables to a random sequence. An independent identically distributed (iid) random sequence is a random sequence X_n in which $\ldots, X_{-2}, X_{-1}, X_0, X_1, X_2, \ldots$ are iid random variables [7]. Similarly, a random sequence is said to be iid if each random variable has the same probability distribution as the other random variables and they are independent of each other. In many real world experiments we observe events repeatedly and when a new observation is an independent realisation of the underlying random phenomena we are dealing with iid random variables. An iid random sequence can be either discrete-value or continuous-value. For the discrete case each random variable X_i has PMF $P_{X_i}(x) = P_X(x)$, while the continuous case, each X_i has PDF $f_{X_i}(x) = f_X(x)$. We start with defining these two classes of random sequences [7].

Theorem 1.4.1. Let X_n denote an iid random sequence. For a discrete-value process, the sample vector $\mathbf{X} = [X(n_1) \dots X(n_k)]'$ has joint PMF

$$P_{\mathbf{X}}(\mathbf{x}) = P_X(x_1)P_X(x_2)\dots P_X(x_k) = \prod_{i=1}^{k} P_X(x_i).$$

For a continuous-value process, the joint PDF of $\mathbf{X} = [X(n_1) \dots X(n_k)]'$ is

$$f_{\mathbf{X}}(\mathbf{x}) = f_X(x_1)f_X(x_2)\dots f_X(x_k) = \prod_{i=1}^{k} f_X(x_i).$$

1.4.5 The Brownian Motion Process

The Brownian motion is a continuous-time, continuous-value stochastic process, also called Wiener process and it is of central importance in stochastic processes theory. Brownian motion describes the movement of a particle in a fluid due to frequent random collisions with the surrounding molecules of the fluid itself [7, 11].

Definition 1.4.7. A **Brownian motion process** $W(t)$ has the property that $W(0) = 0$ and for $\tau > 0$, $W(t+\tau) - W(t)$ is a Gaussian $(0, \sqrt{\alpha\tau})$ random variable that is independent of $W(t')$ for all $t' \leq t$.

For a Brownian motion we can view $W(t)$ as the position of a particle on a line. For small time increment δ, [7]:

$$W(t+\delta) = W(t) + [W(t+\delta) - W(t)]. \tag{1.13}$$

The increment $X = W(t+\delta) - W(t)$ is independent of $W(t)$ and is a Gaussian random variable with mean $\mu = 0$ and variance $\sigma^2 = \sqrt{\alpha\delta}$. This important property of the Brownian motion is called *independent increments* which states that after a time step δ, the position of the particle has moved by X that is independent of the previous position $W(t)$ [7].

Theorem 1.4.2. For the Brownian motion process $W(t)$, the joint PDF of the sample vector $\mathbf{W} = [W(t_1) \dots W(t_k)]'$ is

$$f_{\mathbf{W}}(\mathbf{w}) = \prod_{n=1}^{k} \frac{1}{\sqrt{2\pi\alpha(t_n - t_{n-1})}} e^{-(w_n - w_{n-1})^2/[2\alpha(t_n - t_{n-1})]}.$$

1.4.6 Expected Value and Correlation

As in the case of random variables, stochastic processes are well described using expected values. When dealing with random variables parameters such as the expected value, the variance, covariance and correlation can summarise the information about a probability model. In the case of stochastic processes we deal with functions of time $X(t)$ that provide the corresponding parameters in a straightforward manner and thus encompass complete information about the underlying process. For a stochastic process denoted as $X(t)$, we say that $X(t_1)$ corresponds to the value of the sample function at time t_1 and is a random variable. Therefore it has a PDF $f_{X(t_1)}(x)$ and expected value $\mathbb{E}[X(t)]$. Knowing the PDF everything said about random variables and expected values in previous sections can be applied to $X(t_1)$ and $\mathbb{E}[X(t_1)]$. Since $\mathbb{E}[X(t)]$ is a number, for each value of t the expected value $\mathbb{E}[X(t)]$ is a deterministic function of t. We define the expected value of a process as follows[7]:

Definition 1.4.8 (Expected Value of a Process). The **expected value** of a stochastic process $X(t)$ is the deterministic function

$$\mu_X(t) = \mathbb{E}[X(t)].$$

In general, $\mu_X(t)$ is a function of time and is also called the *ensemble average* of the process $X(t)$.

The covariance function of a stochastic process provides very important information about the time dependence of the process. The covariance $\mathrm{Cov}[X, Y]$ is an indicator of how much information the random variable X provides about random variable Y. When the magnitude of the covariance is high it means that a realisation of X provides an accurate indication about the value of Y. If the two random variables X, Y are observations of $X(t)$ taken at two different times instants, t_1 and $t_2 = t_1 + \tau$, the covariance indicates how much the process is likely to change in the τ instants elapsed between t_1 and t_2. This information is given by the *autocovariance* function [7].

Definition 1.4.9. The **autocovariance** function of the stochastic process $X(t)$ is

$$C_X(t, \tau) = \mathrm{Cov}[X(t), X(t + \tau)].$$

The **autocovariance** function of the random sequence X_n is

$$C_X(m, k) = \mathrm{Cov}[X_m, X_{m+k}].$$

The autocorrelation together with the PDF is considered to contain complete statistical information of a stationary random process [7].

Definition 1.4.10. The **autocorrelation** function of the stochastic process $X(t)$ is

$$R_X(t,\tau) = \mathbb{E}[X(t)X(t+\tau)].$$

The **autocorrelation** function of the random sequence X_n is

$$R_X(m,k) = \mathbb{E}[X_m X_{m+k}].$$

The autocorrelation function of a stochastic process measures the extent of correlation between samples within a random process as a function of how much time elapses between the instants that the samples are taken [7].

Theorem 1.4.3. The autocorrelation and autocovariance functions of a process $X(t)$ satisfy

$$C_X(t,\tau) = R_X(t,\tau) - \mu_X(t)\mu_X(t+\tau).$$

The autocorrelation and autocovariance functions of a random sequence X_n satisfy

$$C_X[n,k] = R_X[n,k] - \mu_X(n)\mu_X(n+k).$$

Next we introduce the concept of stationarity of a stochastic process.

1.4.7 Stationary Processes

Stationarity tells us to what degree the random variables of a stochastic process are constant in time. In a stochastic process $X(t)$, there is a random variable $X(t_1)$ at every instant of time t_1 with PDF $f_{X(t_1)}(x)$. For most random processes, the PDF is time dependent, however there exists a class of random processes known as *stationary processes* where the PDF is time independent, meaning that the statistics of the process is invariant to time shift. Thus, for any two time instants t_1 and $t_1 + \tau$, [7]

$$f_{X(t_1)}(x) = f_{X(t_1+\tau)}(x) = f_X(x). \tag{1.14}$$

For any time shift τ. Eq (1.14) is a necessary but not sufficient condition for a process to be stationary [7].

Definition 1.4.11. A stochastic process $X(t)$ is **stationary** if and only if for all sets of time instants $t_1, \ldots, t_m$ and any time difference τ,

$$f_{X(t_1),\ldots,X(t_m)}(x_1, \ldots, x_m) = f_{X(t_1+\tau),\ldots,X(t_m+\tau)}(x_1, \ldots, x_m)$$

A random sequence X_n is **stationary** if and only if for any set of integer time instants $n_1, \ldots, n_m$ and any time difference k,

$$f_{X_{n_1},\ldots,X_{n_m}}(x_1, \ldots, x_m) = f_{X_{n_1+k},\ldots,X_{n_m+k}}(x_1, \ldots, x_m)$$

The autocovariance and the autocorrelation functions defined in Def 1.4.9 and 1.4.10 are independent of time and depend only on the time shift τ. We adopt the notation $C_X(\tau)$ and $R_X(\tau)$ for the autocovariance function and autocorrelation function, respectively, for a stationary stochastic process [7].

Theorem 1.4.4. For a stationary process $X(t)$, the expected value, autocorrelation and the autocovariance have the following properties for all t:

(i) $\mu_X(t) = \mu_X$

(ii) $R_X(t, \tau) = R_X(0, \tau) = R_X(\tau)$

(iii) $C_X(t, \tau) = R_X(\tau) - \mu_X^2 = C_X(\tau)$.

For a stationary random sequence X_n the expected value, autocorrelation and the autocovariance have the following properties for all n:

(i) $\mathbb{E}[X_n] = \mu_X$

(ii) $R_X[n, k] = R_X[0, k] = R_X[k]$

(iii) $C_X[n, k] = R_X[k] - \mu_X^2 = C_X[k]$.

Presented with the definitions of autocorrelation and autocovariance and their key properties we turn to a more loose form of stationarity of a random process because determining stationarity in the strict sense may be too cumbersome.

1.4.7.1 Wide Sense Stationary Stochastic Processes

There are many applications of probability theory in which we do not have a complete probability model for an experiment. Even so, much can be accomplished with partial information about the model. Often the partial information is included in expected

values, variances, correlations and covariances. In the context of stochastic processes, when these parameters satisfy the conditions of Theorem 1.4.4 we refer to the process as *wide sense stationary* [7].

Definition 1.4.12 (Wide Sense Stationary). $X(t)$ is a **wide sense stationary stochastic process** if and only if for all t,

$$\mathbb{E}[X(t)] = \mu_X, \quad \text{and} \quad R_X(t,\tau) = R_X(0,\tau) = R_X(\tau).$$

X_n is a **wide sense stationary random sequence** if and only if for all n,

$$\mathbb{E}[X_n] = \mu_X, \quad \text{and} \quad R_X[n,k] = R_X[0,k] = R_X[k].$$

Hence a stochastic process $X(t)$ is wide sense stationary if the mean is a constant, μ_X and the autocorrelation depends only on the time shift τ. Theorem 1.4.4 tells us that every stationary process is also wide sense stationary, however, the converse is not necessarily true [7].

1.5 Hypothesis Testing

This section contains a brief introduction to two categories of statistical inference, significance testing and hypothesis testing. Hypothesis testing is part of decision theory based on statistics [7] .

- Significance testing

  Conclusion Accept or reject the hypothesis that the observations result from a certain probability model H_0.

  Accuracy measure Probability of rejecting the hypothesis when it is true.

- Hypothesis testing

  Conclusion The observations result from one of M hypothetical probability models: $H_0, H_1, \ldots, H_{M-1}$.

  Accuracy measure Probability that the conclusion os H_i when the true model is H_j for i, $j = 0, 1, \ldots, M-1$.

1.5.1 Significance Testing

A significance test starts with a null hypothesis, denoted by H_0, that a probability model describes the observations of an experiment. The question addressed by the test has two possible outcomes: accept the null hypothesis or reject it. The *significance level* of the test is defined as the probability of rejecting the hypothesis if it is, in fact, true. The test divides S, the sample space of the experiment, into a space consisting of an acceptance region A and a rejection region $R = A^c$, also called the critical region[6]. If the observation $s \in A$, we accept the null hypothesis, H_0. However, if $s \in R$, we reject it. This decision is compared to a threshold probability, the significance level, denoted by α, and is given by [7].

$$\alpha = P[s \in R] \tag{1.15}$$

We accept only one of the possible hypotheses and reject the other. In doing so we can make two types of errors in our decision [7]:

- **Type I error.** False rejection: Reject H_0 when H_0 is true.
- **Type II error.** False acceptance: Accept H_0 when H_0 is false.

The hypothesis defined in a significance test makes it possible to calculate the probability of a Type I error, given by $\alpha = P[s \in R]$ and is called the *significance level of the test*[6]. In the absence of a probability model for the condition *"H_0 is false"* there is no way to calculate the probability of a Type II error. A *binary hypothesis test*, described in Subsection 1.5.2, includes an *alternative hypothesis* H_1, asides from the null hypothesis. Then it is possible to use the probability model given by H_1 to calculate the probability of a Type II error which is $P[s \in A|H_1]$ [7].

1.5.2 Binary Hypothesis Testing

In a binary hypothesis test we have two possible hypotheses, the null hypothesis denoted by H_0 and the alternative hypothesis denoted H_1 from which only one is assumed to be true. We can have two possible conclusions: *accept* H_0 as the true model, and *accept* H_1. The probability model for the two hypotheses H_0 and H_1 is given by [7]

$$P[H_0] \quad \text{and} \quad P[H_1] = 1 - P[H_0] \tag{1.16}$$

The numbers $P[H_0]$ and $P[H_1]$ are called *a priori probabilities* of the null and alternative hypothesis H_0 and H_1 and give us information about the underlying probability model before we observe an outcome. The test is constructed in a way that it divides the sample space S into two sets, A_0 and $A_1 = A_0^c$. If the outcome $s \in A_0$ then we *accept* H_1. The accuracy measure of the test consists of two error probabilities. $P[A_1|H_0]$ corresponds to the probability of a Type I error, the probability of accepting H_1 when H_0 is the true model. In the same way, $P[A_0|H_1]$ corresponds to the probability of accepting H_0 when H_1 is the actual probability model and leads to a Type II error [7]. The total probability of error in binary hypothesis testing is associated to the a priori probabilities of H_0 and H_1 and to the two conditional probabilities $P[A_1|H_0]$ and $P[A_0|H_1]$ and is given by P_{ERR} [7]:

$$P_{\text{ERR}} = P[A_1|H_0]P[H_0] + P[A_0|H_1]P[H_1]. \tag{1.17}$$

The following theorem specifies the binary hypothesis test that produces the minimum possible total probability error, P_{ERR} [7].

Theorem 1.5.1 (Maximum a posteriori Probability (MAP) Binary Hypothesis Test). Given a binary hypothesis testing experiment with outcome s, the following rule leads to the lowest possible value of P_{ERR}:

$$s \in A_0 \text{ if } P[H_0|s] \geq P[H_1|s]; \qquad s \in A_1 \text{ otherwise.}$$

Theorem 1.5.1 states that in order to minimise P_{ERR} we have to accept the hypothesis with the higher a posteriori probability [7].

1.5.3 Parametric and Nonparametric Methods

Traditional statistical models are based on parametric assumptions such as that data comes from a well known family of distributions for example Gaussian or exponential, where from a set of parameters at least one is unknown and to be inferred trough parametric models, this is possible because the parametric assumptions are met. When this is not possible we have as an alternative the nonparametric models where no previous assumption is made about the underlying distribution. We only require that the data is independent identically distributed from some arbitrary distribution. For this reason nonparametric models are often called distribution-free.

Nonparametric methods require few assumptions about the underlying populations from which the data are obtained. In fact, nonparametric methods forgo the traditional assumption that the underlying populations are normal and are often easier to apply than their normal theory counterparts [12].

1.5.4 Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov (KS) Test is a fully nonparametric test of the equality of continuous, one dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample KS test), or to compare two samples (two-sample KS test), which is what will be used in this case. It allows to test whether two samples come from the same distribution with no need to specify which is the common distribution [3], which is done by evaluating the difference between the two empirical CDFs and denoting the maximum absolute distance as the *KS distance* and denoted by D_{KS}. In Fig. 1.7 the KS distance is marked and was computed by means of Eq. (1.18).

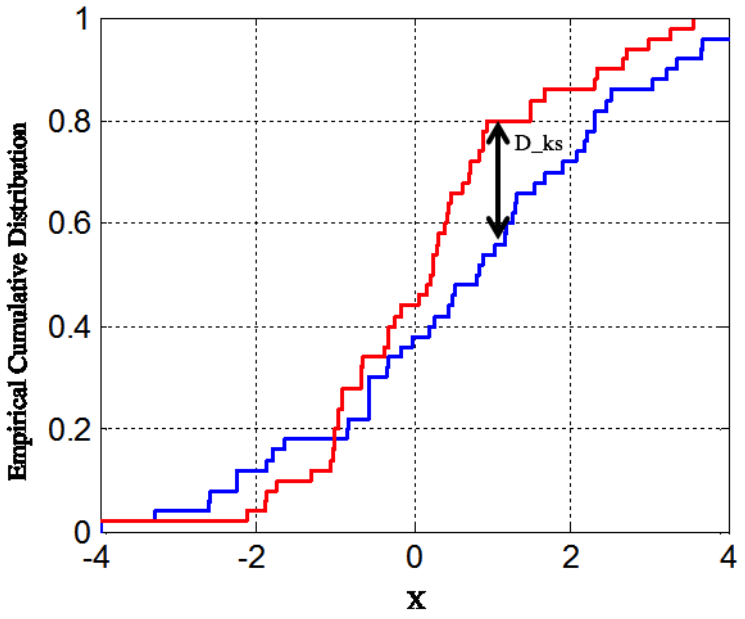


FIGURE 1.7: Two-sample KS test. Two empirical CDFs are plotted and the KS distance, D_{KS}, is located at the largest absolute difference between the two. From [13].

1.5.4.1 Procedure

Essentially we are interested in testing for a common distribution. The simplified procedure for this test is, given the hypothesis H_0: *the two samples come from the same*

distribution, for all x that may be tested against the alternative hypothesis H_1 *the distributions have different empirical cumulative distribution functions (CDFs), for some* x. We compute the CDF for each sample say, $F_1(x)$ and $F_2(x)$, we also calculate and record the difference $F_1(x) - F_2(x)$ [14] . We compare the two sample CDFs by means of the distance; the test statistic is the maximum difference between the two functions.

1.5.4.2 Two Sample Kolmogorov-Smirnov Test

The parameter used in the KS test is the distance between the cumulative distribution functions (CDF) of the two samples and is called the KS distance and denoted as D_{KS} defined below. It is defined as the maximum value of the absolute difference between two CDF. The KS statistic will then be given by [4]

$$D_{KS} = \max_{-\infty<x<+\infty} |F_L(x) - F_R(x)| \tag{1.18}$$

The KS statistic is useful because its distribution in the case of the null hypothesis (data sets drawn from the same distribution) can be calculated, to useful approximation, thus giving the significance of any observed nonzero value of D_{KS}.

The significance level, α of an observed value of D_{KS} (as a disproof of the null hypothesis) can be written in terms on Eq. (1.22) and is given approximately by the formula [4]

$$\alpha = \text{Probability}(D_{KS} > \text{observed}) = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}}\right] D_{KS}\right) \tag{1.19}$$

where N_e is the effective number of data points given by [4]

$$N_e = \frac{N_L N_R}{N_L + N_R} \tag{1.20}$$

for the case of two distributions, where N_L is the number of data points of one data set and N_R is the number in the other data set.[1] The nature of the approximation involved in Eq. (1.19) is that it becomes asymptotically accurate as the N_e becomes large. Also we define the length weighted KS distance, D as follows

$$D = \sqrt{N_e} \cdot D_{KS}. \tag{1.21}$$

[1] The notation of N_R and N_L will become clear after the algorithm is introduced. The idea is that given a time series and a moving pointer, there will occur a cut at some point and the data is broken in two fragments.

The function that enters into the calculation of the significance can be written as [4]

$$Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}e^{-2j^2\lambda^2}. \tag{1.22}$$

1.6 Heavy Tailed Processes

Stochastic models involving heavy tailed distributions becomes more and more popular. Data with so called heavy tails have been collected in fields such as economics, telecommunications, physics, biology (see e.g., [15], [16], [17]). Heavy tailed distributions are probability distributions whose tails are not exponentially bounded, *i.e.*, they have heavier tails than the exponential distribution [18]. Heavy tail also means that there is a larger probability of getting very large values. A particular subclass of these distributions are power-laws, which means that asymptotically the PDF is a power.

We present alternatives of the KS Test which present with more sensitivity in the tails.

1.6.1 Anderson-Darling Test

The Anderson-Darling (AD) test is used to test whether a given sample of data comes from a specific distribution. It is a modification of the Kolmogorov-Smirnov test and gives more weight to the tails than does the KS test.

The statistical problem treated is that of testing the hypothesis that n *i.i.d* random variables have a specified continuous distribution function $F(x)$. If $F_N(x)$ is the empirical cumulative distribution function and $\psi(t)$ is some nonnegative weight function ($0 \leq t \leq 1$), we consider [19]

$$D_{AD} = \max_{-\infty<x<\infty} \sqrt{n}|F_N(x) - F(x)|\sqrt{\psi\left[F(x)\right]}, \tag{1.23}$$

where $\psi(t)$ (≥ 0) is some preassigned weight function and $t = F(x)$. The modification against the KS test is the factor $\sqrt{\psi\left[F(x)\right]}$, the incorporation of a weight function to allow flexibility in the tests. The function $\psi(t)$ is to be chosen as to weight the deviations according to the importance attached to various portions of the distribution function. The selection of $\psi(t) = 1$ yields the criterion of Kolmogorov for Eq. (1.23). The AD distance places more weight on observations in the tails of the distribution with the choice that

$$\psi\left[F(x)\right] = \frac{1}{\left[F(x)(1-F(x))\right]} \tag{1.24}$$

1.6.1.1 M test

The M test is another alternative to the KS test, M test stands for Modified KS test. It is known that the KS test exhibits poor sensitivity to deviations from the hypothesised distribution that occurs in the tails. A modified version of the KS test is introduced that is more sensitive than the original to deviations in the tails. [20].

Let $X_1, \ldots, X_n$ be independent random variables with common continuous distribution F and let $X_{1,n} \leq \ldots \leq X_{n,n}$ denote their order statistics, F_n will denote the continuous empirical distribution function based on $X_1, \ldots, X_n$. Let F_0 be any fixed continuous distribution function [20].

The KS statistic is written as [20]

$$K_n = \max\left\{n^{1/2}\left|F_n(x) - F_0(x)\right| : -\infty < x < \infty\right\}$$

and the Rényi-type statistics as [20]

$$\begin{aligned}
L_{n,1} &= \sup\left\{F_0(x)/F_n(x) : x > X_{1,n}\right\}, \\
L_{n,2} &= \sup\left\{F_n(x)/F_0(x) : -\infty < x < \infty\right\}, \\
U_{n,1} &= \sup\left\{(1-F_0(x))/(1-F_n(x)) : x < X_{n,n}\right\}, \\
U_{n,2} &= \sup\left\{(1-F_n(x))/(1-F_0(x)) : -\infty < x < \infty\right\}.
\end{aligned}$$

Let us consider the following hypothesis test based on the statistics $L_{n,1}, L_{n,2}, U_{n,1}, U_{n,2}$ and K_n for testing [20]

$$H_0 : F = F_0 \quad \text{versus} \quad H_\alpha : F \in \mathcal{F} \quad \text{at level } \alpha,$$

where $\mathcal{F}$ is a specified class of continuous distributions not containing F_0. We reject the null hypothesis H_0 if [20]:

$$\max\left\{w_1 L_{n,1}, w_2 L_{n,2}, K_n, w_3 U_{n,1}, w_4 U_{n,2}\right\} > c$$

where $w_1, \ldots, w_4$ are predetermined nonnegative weight functions and $0 < c < \infty$ is a constant (depending on n) chosen such that the probability of rejection is α. Note that in the case where $w_1 = w_2 = w_3 = w_4 = 0$ this procedure leads to the usual KS test.

Particular versions of the M test are much more sensitive than the KS test to deviations from the hypothesised distribution F_0 in the tails [20].

The finite sample performance of two versions of the M test were examined in [20], one sensitive to light tail alternatives and one sensitive to heavy tail alternatives. These are statistics of the form [20]:

$$L_n = \max\{wL_{n,1}, K_n, wU_{n,1}\} \quad \text{and}$$
$$H_n = \max\{wL_{n,2}, K_n, wU_{n,2}\}.$$

where $w > 0$ is a weight to be specified later [20]. One problem that arises in the practical implementation of these tests is in the determination of criteria for the selection of the weight w.

Numerical evidence indicates that the L_n test is much more sensitive to light tail alternatives than the KS test alone, but less sensitive to heavy tail alternatives; whereas the opposite conclusions are true for H_n. In data analysis it is advised using all three tests which is not very practical [20].

1.7 Segmentation Algorithm

An algorithm for automatically segmenting time series based on differences in CDFs [3], named Kolmogorov-Smirnov segmentation, works as follows. Given a segment of a time series, $\{x_i, i_1 \leq i \leq i_n\}$, a sliding pointer, at $i = i_p$, is moved in order to compare the two fragments $S_L \equiv \{x_{i_1}, \ldots, x_{i_p}\}$ and $S_R \equiv \{x_{i_{p+1}}, \ldots, x_{i_n}\}$, to the left and to the right of i_p. For each i, the KS statistic $D_{KS}(i)$ and the value $D(i)$ is computed. The position i_p of the pointer is moved such that the sizes of the two segments are at least unitary. Then one selects the position $i_{\max}$ that maximizes the KS statistic given by Eq. (1.18), between the two patches on the left and on the right of i_p. Once found, the position $i_{\max}$ of the maximum distance D, $D_{\max}$, we check where the statistical significance of a potentially relevant cut at $i_{\max}$ by comparison with the result that would be obtained was the sequence random [5]. If we compared two random samples that are statistically independent, this would be Q_{KS}. Since we compare the maximum

of two positions at i in a time series, that are statistically not independent, this must be compared to $D_{\max}^{\text{crit}}$ criterion given by Eq. (2.3) further detailed in Chapter 2. For a potential cut is then checked if $D_{\max}$ exceeds its critical value $D_{\max}^{\text{crit}}$, for the selected significance level. Before final acceptance of the cut, we can still require a minimum size l_0, namely $i_{\max} - i_1 + 1,\ i_n - i_{\max} \geq l_0$. The procedure is then recursively applied starting off from the entire series $\{x_i,\ 1 \leq i \leq N\}$, where N is the total number of data points, until no significant cuts are found within some segment (See Appendix A).

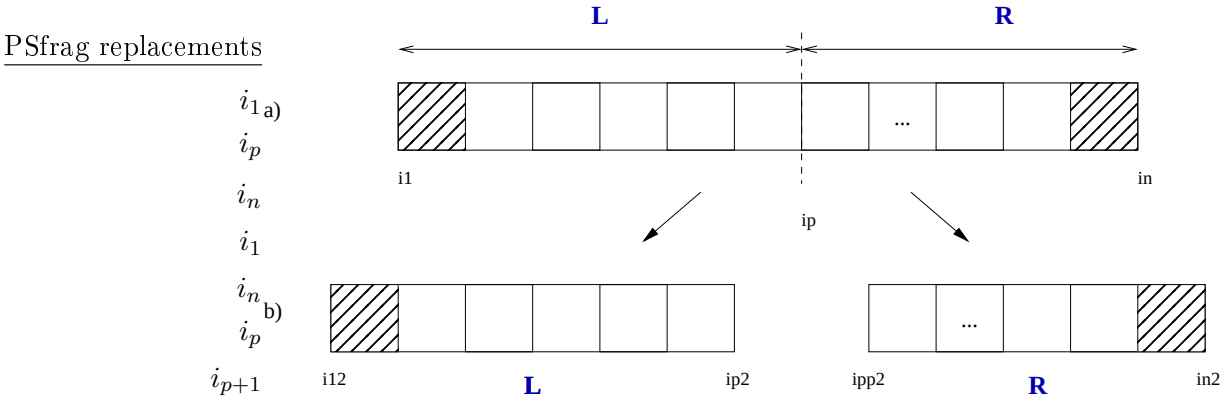


FIGURE 1.8: Ilustration of the segmentation algorithm for one iteration.

In Fig. 1.8 we see an illustration of the described segmentation algorithm for just one iteration. It shows a time series of length n fragmenting at the position where the pointer is at, i_p which results in two segments, one on the left (L) and one on the right (R) of i_p.

Chapter 2

Nonparametric Segmentation of Nonlinear and Heavy Tailed Time Series

In this section we perform extensive numerical tests using the segmentation algorithm explained in Section 1.7 based on the two sample KS test, described in Subsection 1.5.4 and [3]. In a first scenario we undertake basic testing on artificial time series generated from Gaussian distributions to understand how the KS test behaves when presented with time series composed of data coming either from the same or from different distributions, we want to explore how accurate the segmentation algorithm is at finding these differences, if existing, within a generated Gaussian time series. The classic KS test assumes that the tested data are independent, however this is not our case because we run the test on an entire time series and the results are interdependent fragments of this time series. For this reason we shift our attention to an empirical statistical significance criterion that permits interdependence between the fragments to be tested [3] and is therefore more accurate in our case. In Section 2.4 we evaluate the accuracy of this new significance criterion for the normal distribution and non-normal distributions such as log-normal and Cauchy, which differ strongly from Gaussian in tail behaviour. The classical KS test suffers from a flaw, the test is weakly sensitive in the tails of the tested sample, when it is often these tail events that one is most interested in [21]. For this reason we introduce a modification to the segmentation algorithm, replacing the KS test with the AD test [19], incorporating a weight function to allow more flexibility in the test. This weight function is chosen such that it accounts for the tails. The described modification is known as the Anderson-Darling test described in Section 1.6.1. We are interested in how much better does the AD test work on distributions with tail behaviour rather than the KS test. Afterwards investigate the efficiency and performance range

of the KS segmentation algorithm for each of the distributions (Gaussian, log-normal and Cauchy) to establish a comparison. We then test the segmentation algorithm with the AD implementation on a heavy tailed distribution, the Cauchy distribution. When studying hypothesis tests that assume normality, seeing how the tests perform on data from a Cauchy distribution is a good indicator of how sensitive the test is to heavy-tail departures from normality.

2.1 Significance

We start with plotting the classic KS test in Fig. 2.1 given by the probability function Q_{KS} from Eq. (1.22) in function of λ which corresponds to the argument in brackets of Eq. (1.19) shown in Eq. (2.1), in the large limit approximation, $N_e \to \infty$ (Note that Eq. (1.22) is defined as an infinite sum and for in this numerical test has been truncated to size 100.) [3]:

$$\lambda = \left(\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}} \right) D_{KS} \tag{2.1}$$

From Eq. (2.1) we see that for large N_e, λ is approximately equal to $D_{KS} \cdot \sqrt{N_e}$. When $N \to \infty$ the distribution of the weighted KS statistic, $D = D_{KS} \cdot \sqrt{N_e}$ is asymptotically the KS distribution with CDF given by $1 - Q_{KS}$. Shown in Fig. 2.1 is the probability function Q_{KS} given by Eq. (1.22) overlapped with three points that correspond to $Q_{KS}(\lambda)$ values equal to 0.1, 0.01 and 0.05 versus the critical values [13] for λ equal to 1.22, 1.36 and 1.63, [13], respectively, for the large N_e limit, $N_e \to \infty$. Critical values are defined as the threshold value delimiting the regions of acceptance and rejection for the test statistic.

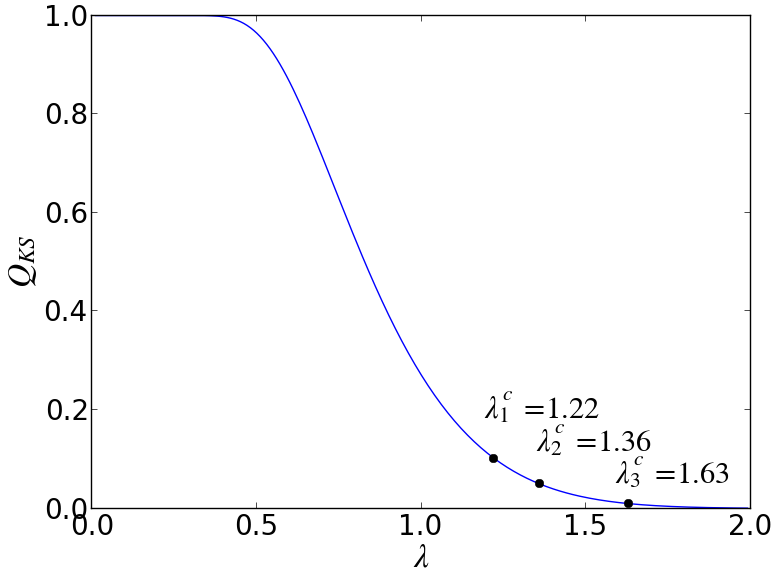


FIGURE 2.1: KS probability function Q_{KS} as a function of λ in the limit of large sample sizes, $N \to \infty$, for which λ is approximately equal to $D_{KS} \cdot \sqrt{N_e} \approx D$. When $N \to \infty$ the distribution of the weighted KS statistic, $D = D_{KS} \cdot \sqrt{N_e}$ is asymptotically the KS distribution with CDF given by $1 - Q_{KS}$. Plotted is the probability function Q_{KS} given by Eq. (1.22) overlapped with three points that correspond to Q_{KS} values equal to 0.1, 0.01 and 0.05 versus the critical values for λ, λ^c, [13] 1.22, 1.36 and 1.63, respectively, for the large N_e limit, $N_e \to \infty$. These critical points fall perfectly onto the curve of the probability function Q_{KS}.

The three points in Fig. 2.1 correspond to large N_e limit values for Q_{KS} equal to 0.1, 0.01 and 0.05 and their respective critical λ values, corresponding to λ^c equal to 1.22, 1.36 and 1.63, [13], which mark the values of the inverse of the KS probability function, Q_{KS}^{-1}, where we can distinguish time series from different distributions with 90%, 95% and 99% confidence levels, respectively.

Interpreting Fig. 2.1 we remark that for higher λ the risk of wrongly discarding the null hypothesis is low meaning that the probability that we make an error in saying that they are different becomes very low. In this case we can say that it is very likely that the distributions are not the same. On the other hand for small values of λ the significance is very high hence we are dealing with a high risk and cannot safely reject the null hypothesis that the distributions are the same. For example for $\lambda^c = 1.22$, only in 10% of the cases we would expect values of $\left(\sqrt{N_e} + 0.12 + \frac{0.11}{\sqrt{N_e}}\right) D_{KS}$ to be greater than 1.22. This means that we can discard the null hypothesis of equality of distributions with a 90% confidence.

From this we see that Eq. (1.22) is accurate in computing the significance for infinite sized time series.

Next, taking into account that $D = D_{KS} \cdot \sqrt{N_e}$ we substitute D_{KS} and in Eq. (1.19) obtain

$$\text{Prob}(D > \text{observed}) = Q_{KS}\left(\left[1 + \frac{0.12}{\sqrt{N_e}} + \frac{0.11}{N_e}\right] D\right) \tag{2.2}$$

We want to plot the inverse of Q_{KS} which is a linear function of D. We take the inverse of Eq. (2.2) and obtain Fig. 2.2.

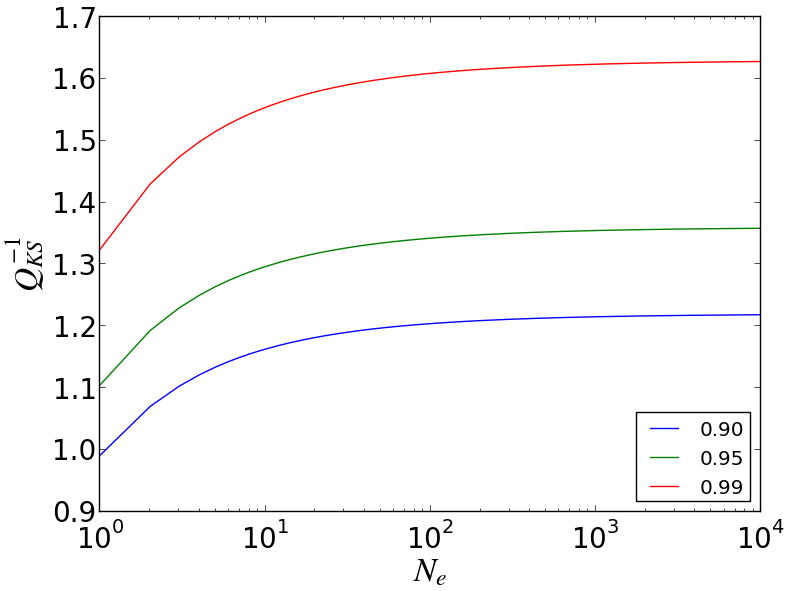


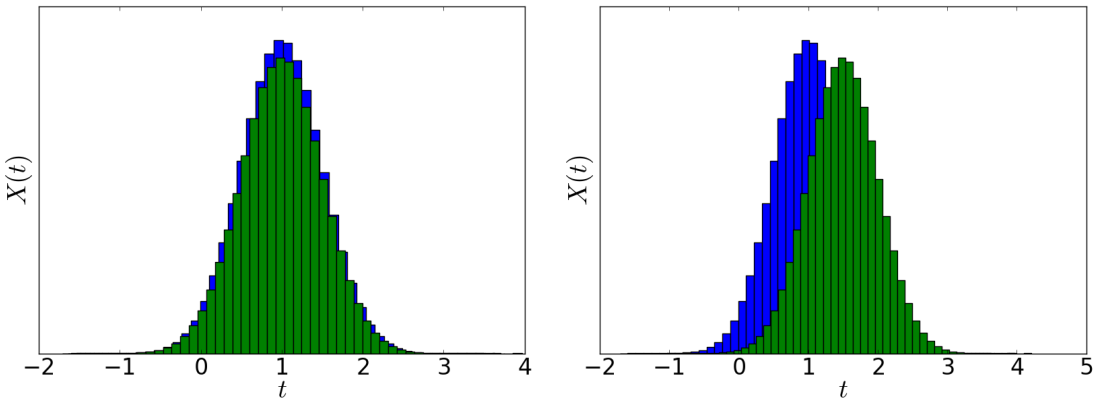
FIGURE 2.2: Inverse of the probability function, Q_{KS} as a function of the time series length N_e. Three confidence levels are shown, 90%, 95% and 99%, corresponding to significance levels α of 0.1, 0.05 and 0.01, respectively, for the classical KS test, given by the inverse of Eq. (2.2). The curves asymptotically tend to the critical values, 1.22, 1.36 and 1.63, [13], from top to bottom.

The Q_{KS} criterion, given by Eq. (1.22), is derived under the assumption that we have independent time series. We see that for each significance level the Q_{KS} is a monotonically increasing function of sample size N_e and asymptotically attains the critical values 1.22, 1.36 and 1.63.

In our case, however, where we iteratively segment a time series, the segments are not independent, we have a finite time series that is divided into fragments, for this reason we need a better significance criterion. This criterion is given by Eq. (2.3), [3].

2.2 Kolmogorov-Smirnov Performance Test

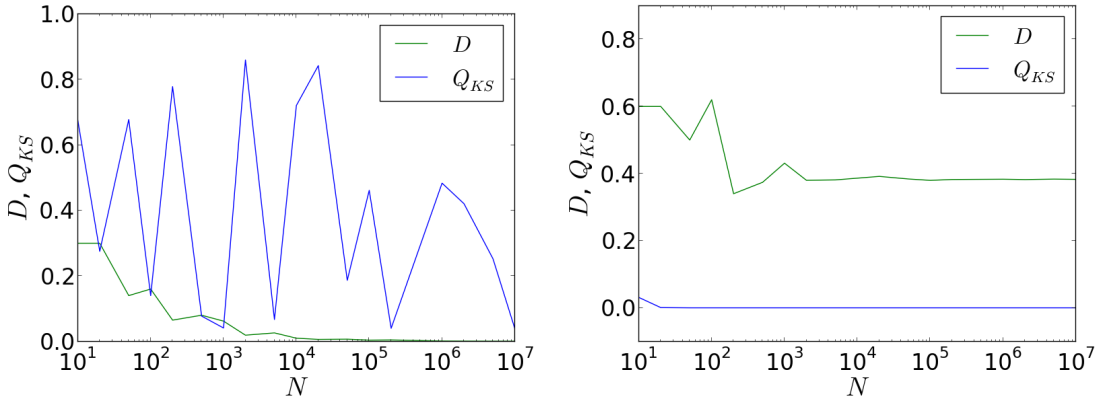
We now evaluate the ability of the KS test to determine whether two samples are drawn from the same distribution or not. For this we generate two samples of Gaussian random numbers and compute the KS statistic, D_{KS}, and the probability function Q_{KS} of both samples. In order to simulate time series composed of two segments with Gaussian distributions, obtained when the distribution parameters are the same in both segments, Fig. 2.3 (A), and in Fig. 2.3 (B) we simulate the case for a sample where data is mixed of two distributions choosing different distribution parameters. In the first case are generated two samples of Gaussian random numbers with identical parameters, same mean value $\mu_1 = \mu_2 = 1.0$ and standard deviation, $\sigma = 0.5$. The corresponding histograms are plotted in Fig. 2.3 a). For the second case we draw two samples but this time with different means, $\mu_1 = 1.0$ and $\mu_2 = 1.5$, while the standard deviation remains unchanged. This is illustrated in Fig. 2.3 b) where we can see that the histograms are shifted representing two samples from different distributions. This test makes use of the implemented two-sample KS test in the Statistical functions of Python scipy package which computes the Kolmogorov-Smirnov statistic on two samples. It tests the rejection of the null hypothesis that samples are drawn from the same distribution. If the K-S statistic is small or the probability function Q_{KS} value is high, then we cannot reject the null hypothesis that the distributions of the two samples are the same [22].



(A) (B)

FIGURE 2.3: PDF of two random samples generated from different Gaussian distributions on which we make the performance test of the two sample KS test, (A) has two samples with same mean values and (B) has mean values shifted by 0.5. Histograms of two samples from a Gaussian distribution with (A) means $\mu_1 = \mu_2 = 1.0$ and (B) means $\mu_1 = 1.0$, $\mu_2 = 1.5$

Given two time series each of length N we plot the KS probability function, Q_{KS}, and the KS distance, D against N.



(A) (B)

FIGURE 2.4: Evaluation of the performance of the two sample KS test in determining whether two samples are drawn from the same distribution. Plot of KS probability function, Q_{KS}, and the length-weighted distance between CDFs, D as functions of the size of the samples, N. In (A) the KS probability function is oscillating discontinuously with growing N and the distance decreases to zero, in this case we cannot safely reject the null hypothesis that they are from the same distribution. On the other hand, in (B) for both quantities we get a good result, the KS probability function rapidly falls to zero for very short time series and the KS distance also decreases to zero from where we can exclude the null hypothesis even for short time series.

We know that if the KS statistic D_{KS} is small or the probability function Q_{KS} is high we cannot reject the hypothesis that the distributions of the two samples are the same. On the left plot of Fig. 2.4 the probability value Q_{KS} seems to oscillate discontinuously with growing N, even for very large N, whereas the KS distance D decreases monotonically as expected. The probability of wrongly rejecting the null hypothesis, that they are the same, is so high that we cannot actually reject it. The obtained result for Fig. 2.4 (A) show us that this is not sufficient to infer that they belong to the same distribution or not, which is a known problem, see for example Chapter 14.3 in [4].

Next we make the same test but this time with two different distributions, as shown in Fig. 2.3 (B), the result for this case is shown in in Fig. 2.4 (B).

The KS statistic D tends to a stable value, zero, which is a positive result. For the probability function Q_{KS} we get a perfect result, the curve rapidly tends to zero translating very low probability values of Q_{KS}. Hence even for short time series of different distributions the null hypothesis is already completely excluded.

The two sample KS test can reject the null hypothesis that the two distributions are the same if there is a sufficiently high significance , Fig. 2.4 (B), and a large enough length N. However, this does not imply that the test is capable of deciding the opposite question, whether the two distributions are the same, Fig. 2.4 (A).

The distance must tend to a stable value because the curves become finer and finer as N grows and if they come from the same distribution they need to converge to the same curve which is the underlying CDF. In this case there is no difference between the CDF's and the distance between them vanishes. On the other hand if they do not come from the same distribution the distance tends to some limiting value.

Now that we studied the ability of the KS test to detect whether two samples are drawn from the same distribution we are motivated to look further into the performance of the KS segmentation algorithm. In the next Section we evaluate the accuracy of the algorithm in detecting where known differences in distribution parameters exist within an artificial time series.

2.3 A KS Segmentation Algorithm for Nonstationary Time Series

2.3.1 Testing Segmentation of Time Series of Random Samples from a Gaussian Distribution

In this section we perform numerical tests to evaluate the accuracy of the algorithm described in Section 1.7 in detecting differences within an artificial time series where positions of expected segmentation are known. For this we run the algorithm once through a compound Gaussian time series formed by one or two sets of random numbers each. Four time series of total size $N = 200$ where $N = N_A + N_B$ are generated in the following way, Data (A) has the whole time series with parameter values, mean $\mu = 1$ and standard deviation $\sigma = 0.5$ corresponding to a case where there is no difference within the time series thus we expect that the algorithm does not find any relevant cut position. Data (B) and Data (C) are time series composed of two patches of sizes $N_A = 100$ and $N_B = 100$, patch A has parameters mean $\mu = 1$ and standard deviation $\sigma = 0.5$. In case (B) patch N_B has mean $\mu = 1.5$ and standard deviation $\sigma = 0.5$, in case (C) patch N_B has smaller mean, $\mu = 1.1$ and the standard deviation is the same as in data (B), $\sigma = 0.5$. These two cases illustrate time series composed of different distributions and we expect the segmentation algorithm to find a relevant cut at position $N = 100$ because this is where there is a change in parameters and hence difference in distributions. Finally, case (D) has patch sizes $N_A = 30$ and $N_B = 170$, patch N_A has the same parameters as before, mean $\mu = 1.0$ and standard deviation $\sigma = 0.5$, while patch N_B has mean $\mu = 1.5$ and standard deviation $\sigma = 0.5$, therefore we expect the algorithm to cut at position $N = 30$. This final test on case (D) serves to see how sensitive the test is when applied to shorter segments.

TABLE 2.1: Parameters of the generated time series composed of two sets of numerically drawn Gaussian random numbers in order to obtain different distributions within a single time series.. Each time series, (A), ..., (D), is composed of two patches, A and B, each of size N_A and N_B with total size $N = N_A + N_B = 200$. First three columns correspond to the parameters for patch A while the remaining three columns describe patch B, where are specified the respective means $\mu_{A,B}$, the standard deviations $\sigma_{A,B}$ and the patch sizes $N_{A,B}$.

	μ_A	σ_A	N_A	μ_B	σ_B	N_B
Data (A)	1	0.5	200			
Data (B)	1	0.5	100	1.5	0.5	100
Data (C)	1	0.5	100	1.1	0.5	100
Data (D)	1	0.5	30	1.5	0.5	170

Given these generated time series we executed a single iteration of the segmentation algorithm described in Section 1.7 and plotted the position i against the maximum distance D_{KS} given by Eq. (1.18), D which is simply the distance between the CDFs at each position, calculated with $D(i) = \sqrt{N_e} D_{KS}(i)$, the $D_{\max}^{\text{crit}}$ significance criterion given by Eq. (2.3) and the square root of the effective number of points, N_e, Eq. (1.20). The plots for each case are shown in Fig. 2.5. For this simulation was introduced a significance criterion, according to [3], given by the heuristic simple expression, where $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ for $P_0 = 0.90$, 0.95, and 0.99, respectively [3].

$$D_{\max}^{\text{crit}}(N_e) = a(\ln N_e - b)^c. \tag{2.3}$$

For this simulation we chose the parameter values corresponding to a confidence level of $P_0 = 0.95$.

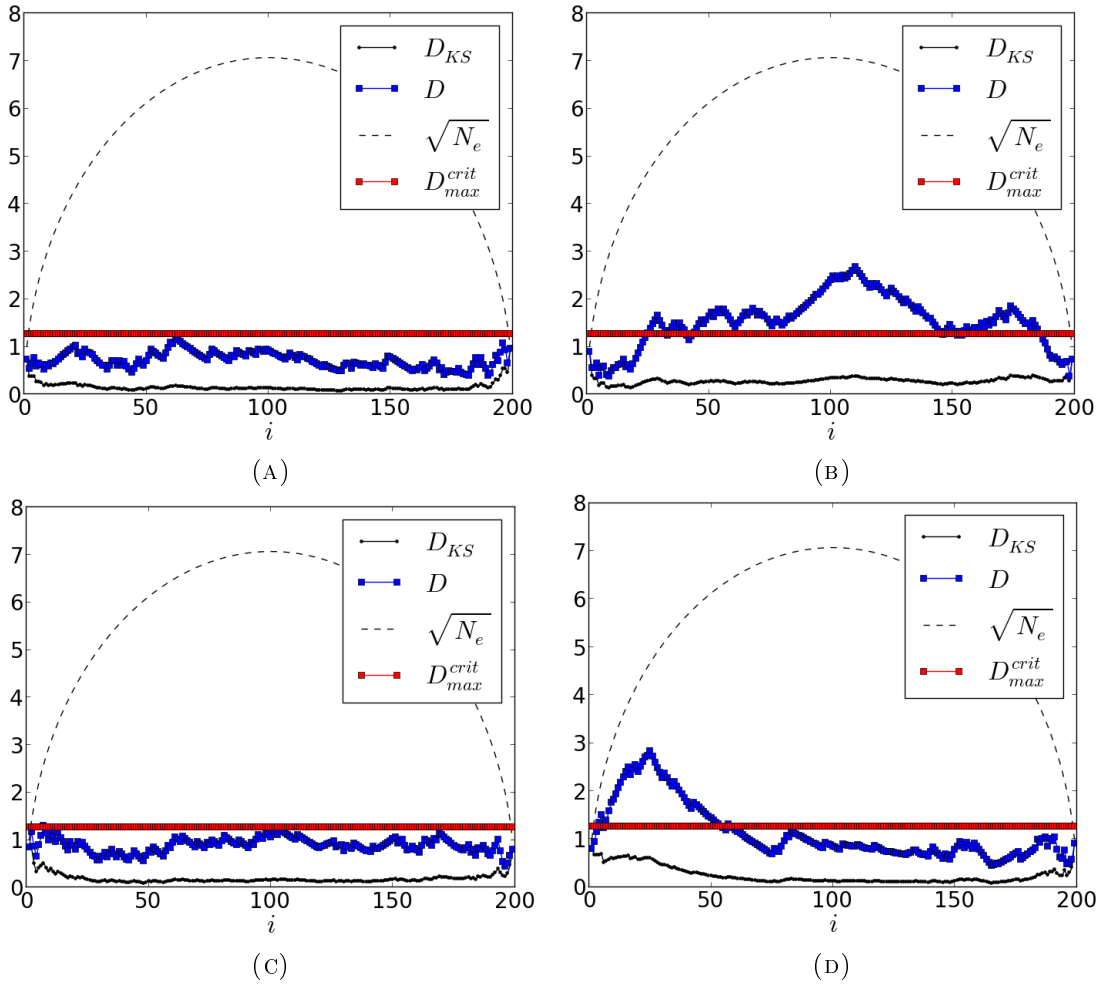


FIGURE 2.5: Accuracy of the KS segmentation algorithm for compound time series created from pairs of Gaussian distributions. The KS segmentation algorithm on non-stationary time series and the relevant quantities are shown. This evaluates the accuracy of the algorithm to detect diferent distributions inside some time series. Time series are composed of one or two segments A, B with parameters μ, σ and length $N_{A,B}$ given by Table 2.1 composed of generated random samples from a Gaussian distribution. Data (A) corresponds to a gaussian with mean 1 and standard deviation 0.5, while data (B) and data (C) differ in the right half of the time series (patch B) at their mean values of 1.5 and 1.1 respectively. Data (D) has one patch of size 30 and the other patch with size 170 with mean 1 and 1.5 and equal standard deviation 0.5. For each of the four cases we run the algorithm once and for each we plot the curves for D_{KS}, D, $D_{\max}^{\text{crit}}$ and $\sqrt{N_e}$. We can see that our segmentation algorithm cuts exactly where it should, according to the acceptance criterion described in Section 1.7, at the point where the distributions change from time series A to B, for plots (B) and (D), or not at all as expected for (A) and finally for (C) the parameter difference is too subtle for the cut to be significant.

As said in Section 1.7, our cut acceptance criterion is that if $D_{\max}$ exceeds its critical value $D_{\max}^{\text{crit}}$ the cut is accepted, this is illustrated in Fig. 2.5a by the blue and red lines, respectively. If we look for the maximum of the blue line we find our $D_{\max}$ and if this point is above the red line, $D_{\max}^{\text{crit}}$, then it should cut at that point.

For Data (A) we obtain Fig. 2.5a where the distance between CDFs D, corresponding to the blue line, is always below the red line, the critical $D_{\max}$ with 95% significance, this means that is makes no sense to cut this time series at any position i, because the time series is composed of samples belonging to the same distribution.

The other plots, however, contain samples from different distributions, hence we expect that the algorithm cuts the time series at some point. Data of Fig. 2.5b tells us clearly that the data belongs to different distributions and also that the position $i_{\max}$ at which there will be a cut corresponds to the position $i = 100$. This is expected because the sample is made of two patches each with 100 random numbers, one with mean value 1.0 and the other with mean value 1.5 while standard deviation is the same, recalling Table 2.1 . For Fig. 2.5c we put the mean value of the second half, $i > 100$, to 1.1. This case it is more subtle that the previous, we can see that the blue line never actually exceeds the red line but has its maximum value at approximately $i = 110$, where the algorithm should perform the segmentation of the time series. For the last case, Data (D), mean value and standard deviation are the same as for (a) changing now the size of each patch to $N_A = 30$ and $N_B = 170$. Again we see that the position where D has its maximum value is at approximately $i = 30$, which is exactly what we expected.
D_{KS} has a U shape because at the ends, $i \to \{0, N\}$, we compare empirical CDFs of one very short and one very long time serie. $\sqrt{N_e}$ has a inverse U shape because of the limits $N_L \to 0$, $N_R \to N$, $N_R \to 0$, $N_L \to N$, $N_R \to N_L \to \frac{N}{2}$ in Eq. (1.20). The ends of the time series have effects on D_{KS} and N_e that compensate each other, which means we do not overemphatize the end points. D has, however, an almost flat shape with fluctuations because of the product $D = D_{KS} \cdot \sqrt{N_e}$. $D_{\max}^{\text{crit}}$ is a horizontal line because of its expression, given by Eq. (2.3), that depends only on the constant parameters a, b, c and the length N_e.

The positive results of this test indicate us that the segmentation algorithm described in Section 1.7 is accurate at finding differences within a time series, this motivates us to perform further tests on distributions other than the Gaussian to evaluate the accuracy of the algorithm when working with non-normal distributions like log-normal and Cauchy. In the next section we study the statistical significance criterion to know how likely is it for a given finite stationary time series of length N to find values D_{KS} greater than a certain threshold $D_{\max}^{\text{crit}}$, for a given significance level, and what is the distribution of these values.

2.4 Statistical Significance Criterion

In this section we explore a new significance criterion for the segmentation algorithm, suggested by [3], different than the Q_{KS} criterion for the two-sample KS test. The Q_{KS} criterion assumes independence between tested fragments while in our case we segment one time series in a number of small fragments which are interdependent. We want to know how likely is it for a given finite stationary time series of length N to find values for the KS distance D_{KS} greater than a certain threshold, the cut acceptance criterion for the KS segmentation algorithm $D_{\max}^{\text{crit}}$, for a given significance level, and what is the distribution of these values for Gaussian, log-normal and Cauchy distributions. Also we implemented a modification to the segmentation algorithm using the Anderson-Darling Test, described in Subsection 1.6.1, which adds a weight term in order to account for tail behaviour of a distribution, and obtained new parameter values (a, b, c) for Eq. (2.3).

2.4.1 Description

We generate $R = 20000$ sequences of different lengths, N, from 10 up to 40000. For each of these $R \times N$ trials we perform a single iteration of the segmentation algorithm described in Subsection 1.7, i.e, we run the pointer through the time series and calculate the KS statistic D_{KS} at each position of the pointer, calculate the maximum distance D_{max} which is the maximum for all positions i of the pointer over the D_{KS} values.

We thus obtain for each length N a series of R trial values, which gives us the distribution of the $D_{\max}$ values that would be expected if there is a homogeneous distribution. We can derive from this distribution a criterion that allows us to reject the null hypothesis of single distribution at a significance level P_0.

2.4.2 Numerical Results

In order to obtain the critical curves for significance testing the maximal distance, $D_{\max}$, is determined numerically for a large number (20000) of sequences of N random numbers generated from Gaussian, log-normal and Cauchy distributions. We obtain the critical values of $D_{\max}(N)$, $D_{\max}^{\text{crit}}(N)$, for each confidence level P_0, 90%, 95% and 99%.

2.4.2.1 Gaussian Distribution

First we want to generate the critical values of $D_{\max}$, [3], for a Gaussian distribution according to the process described in Subsection 2.4.1. Running the KS algorithm once

over each of the time series we obtain a distribution of values of $D_{\max}$ for each confidence level. For Fig. 2.6 we plot the critical curves determined for a Gaussian distribution with standard deviation $\sigma = 0.5$ and zero mean.

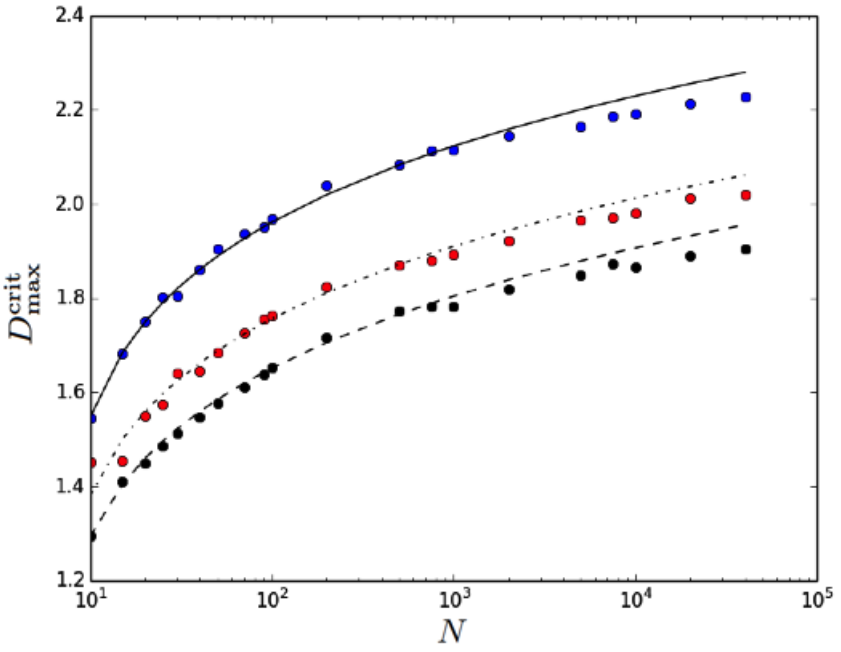


FIGURE 2.6: Critical curves for significance testing determined for a Gaussian distribution. Critical values of the maximal distance,$D_{\max}$, $D_{\max}^{\text{crit}}$, as a function of the sequence length N, up to 40000, for series of numbers generated from a Gaussian distribution with standard deviation $\sigma = 0.5$ and mean $\mu = 0$. The lines correspond to Eq. (2.3) at different confidence levels P_0, 90%, 95% and 99%, from bottom to top. The obtained percentile values of the critical values of $D_{\max}$ follow the same trend as the curves.

The points in Fig. 2.6 correspond to the distribution of the critical values of $D_{\max}$ for a Gaussian distribution while the three curves are plotted from Eq. (2.3) for different confidence levels P_0, 90%, 95% and 99%, from bottom to top. The numerically obtained points follow the same trend than the curves, there is no overlap and we observe that they grow monotonically which represents a different behaviour than the asymptotic one observed in Fig. 2.2. We conclude that Eq. (2.3) is a good fit for our obtained data. The parameters used in Eq. (2.3) are $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ from [3], at different confidence levels, 90%, 95% and 99%, respectively, from bottom to top.

Now that we obtained the distribution of the critical values of $D_{\max}$ for a Gaussian distribution we want to turn our attention to other classes of distributions, namely log-normal and Cauchy, which present tail behaviour.

2.4.2.2 Log-normal Distribution

We want to explore the sensitivity of the significance criterion given by Eq. (2.3) for a distribution with tail characteristics different from a Gaussian distribution, in this following case, the log-normal distribution. We proceed as detailed in Subsection 2.4.1 but the $R = 20000$ time series are now generated from a log-normal distribution described in Subsection 1.2.2.4 with scale parameter $\sigma_Y = 1$ and location parameter $\mu_Y = 0$, according to Eq. (1.2.14). The result of running the KS algorithm once over each of the time series are the percentile values for the critical values of $D_{\max}$ at different confidence levels, 90%, 95% and 99%, that is shown in Fig. 2.7.

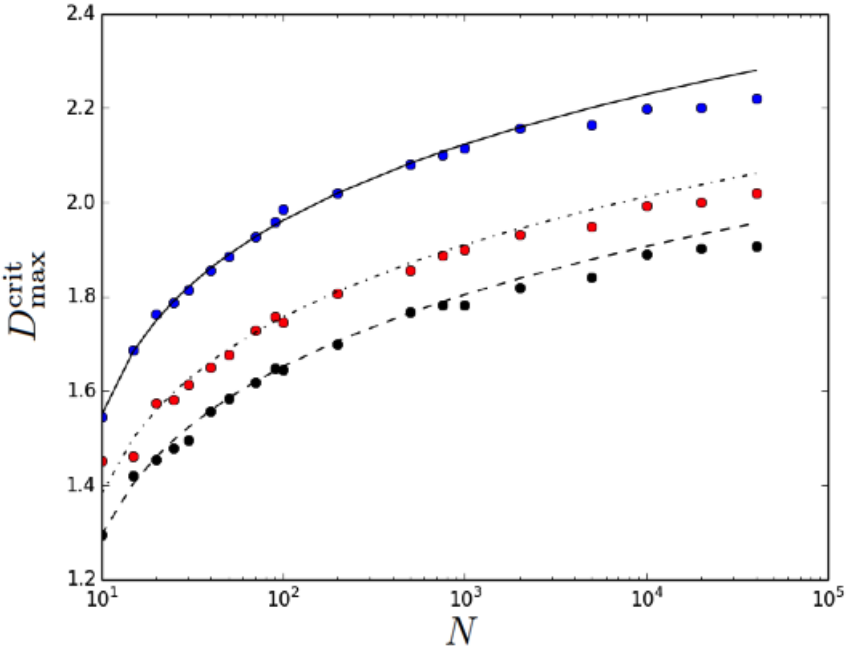


FIGURE 2.7: Critical curves for significance testing determined for a log-normal distribution. Critical values of of the maximal distance, $D_{\max}$, $D_{\max}^{\text{crit}}$, as a function of the sequence length N, up to 40000, for series of numbers generated from a Log-normal distribution with scale parameter $\sigma_Y = 1$ and location parameter $\mu_Y = 0$. The curves correspond to Eq. (2.3) with $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ from [3], at different confidence levels, 90%, 95% and 99%, from bottom to top. The obtained percentile values of the critical values of $D_{\max}$ follow the same trend as the curves.

The points in Fig. 2.7 correspond to the distribution of the critical values of $D_{\max}$ for a log-normal distribution while the three curves are plotted from Eq. (2.3) for different confidence levels. We can see a similarity with the Gaussian case of Subsection 2.4.2.1 shown in Fig. 2.6. The curves fall on the curves hence the fit given by Eq. (2.3) with parameters $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ at different confidence levels, 90%, 95% and 99%, respectively, is a striking good one although the

underlying distributions have completely different characteristics. Again, no overlap is observed between data and the data points follow a monotonical growth

2.4.2.3 Cauchy Distribution

The log-normal distribution has tails heavier than the normal distribution and still our results show that Eq. (2.3) is a good fit. Now we choose a distribution with power law tails, the Cauchy distribution which is widely used in physics, described in Section 1.2.2.5, and evaluate if the fit is still good. This distribution is an example of a pathological distribution since its mean and variance are undefined and does not have finite moments of order greater or equal to one nor does it have a moment generating function [23].

Recalling the probability distribution for Cauchy from Section 1.2.2.5 and setting the location parameter $x_0 = 0$ is then given by Eq. (1.10)

$$f_X(x) = \frac{1}{\pi\gamma\left[1 + \left(\frac{x}{\gamma}\right)^2\right]}. \tag{2.4}$$

Running the KS algorithm once over each of the time series results in a distribution for the critical values of $D_{\max}$ at different confidence levels, 90%, 95% and 99%, that is shown in Fig. 2.8.

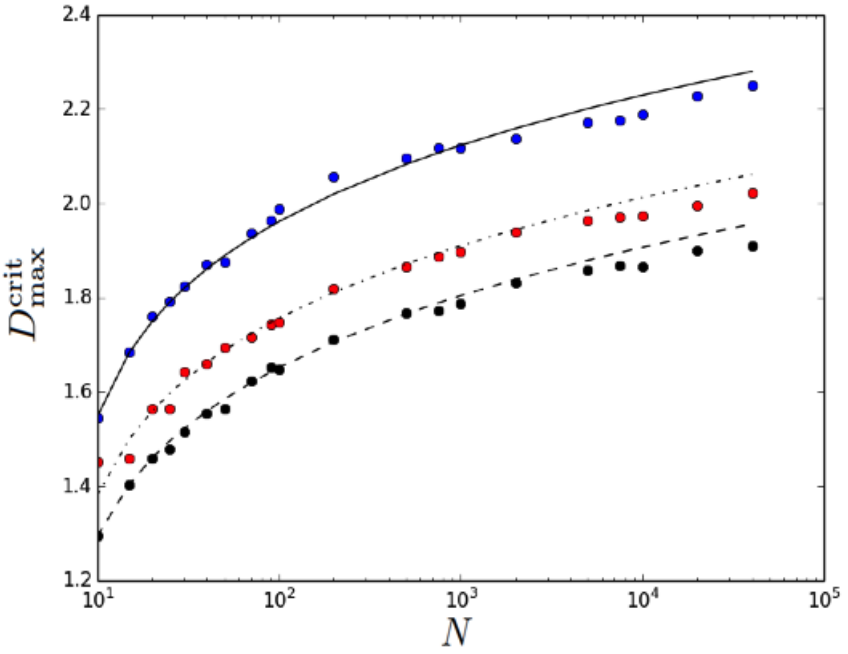


FIGURE 2.8: Critical curves for significance testing determined for a Cauchy distribution. Critical values of $D_{\max}$ as a function of the sequence length N, up to 40000, for series of numbers generated from a Cauchy distribution with $\gamma = 1$. The lines correspond to Eq. (2.3) at different confidence levels, 90%, 95% and 99% from bottom to top. The obtained percentile values of the critical values of $D_{\max}$ follow the same trend as the curves.

In Fig. 2.8 we have the percentile values for the critical values of the acceptance criterion, $D_{\max}^{\text{crit}}$, as a function of the sequence length N.

We generate time series of random numbers from a Cauchy distribution with scale parameter $\gamma = 1$ for a sequence length N with a maximum of 40000, averaged over R realisations. The lines correspond to fits using Eq. (2.3) with parameters $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ at different confidence levels, 90%, 95% and 99%, respectively.

From Fig. 2.8 it is clear that the results are a good fit and follow the tendency of the lines. We also see that the curves are monotonically growing and that there is no overlap between the three and again quite similar to the previous cases for Gaussian distribution shown in Fig. 2.6 and log-normal shown in Fig. 2.7.

The numerical results shown in Figs. 2.6, 2.7 and 2.8 reveal that the significance criterion given by Eq. (2.3) is applicable for samples from normal and non-normal distributions with tail behaviour. We now have the $D_{\max}^{\text{crit}}$ values to decide whether a cut in time series composed of Gaussian, log-normal or Cauchy distributions makes sense at a given significance.

With these results we switch our attention to the AD test to seek for improvements.

2.4.2.4 Anderson-Darling Test for Gaussian Distribution

The Anderson-Darling (AD) test is a modification of the KS test and gives more weight to the tails than does the KS test, described in Subsection 1.6.1. For this case we modified the segmentation algorithm to use in each iteration the AD test instead of a KS test, where instead of just calculating the absolute difference D_{KS} between two points of the empirical CDF we add a weight term. We consider [19],

$$D_{AD} = \max_{-\infty<x<\infty} \sqrt{N_e}|F_L(x) - F_R(x)|\sqrt{\psi\left[F(x)\right]}, \tag{2.5}$$

The modification against the KS test is the factor $\sqrt{\psi\left[F(x)\right]}$. The AD distance gives, in this way, more weight to the tails of the distribution with the choice that

$$\psi(t) = \frac{1}{[t(1-t)]}, and \tag{2.6}$$

since we compare two empirical distribution functions, $F_L(x)$ and $F_R(x)$, we chose t to be the mean $1/2$ $(F_L(x) - F_R(x))$.

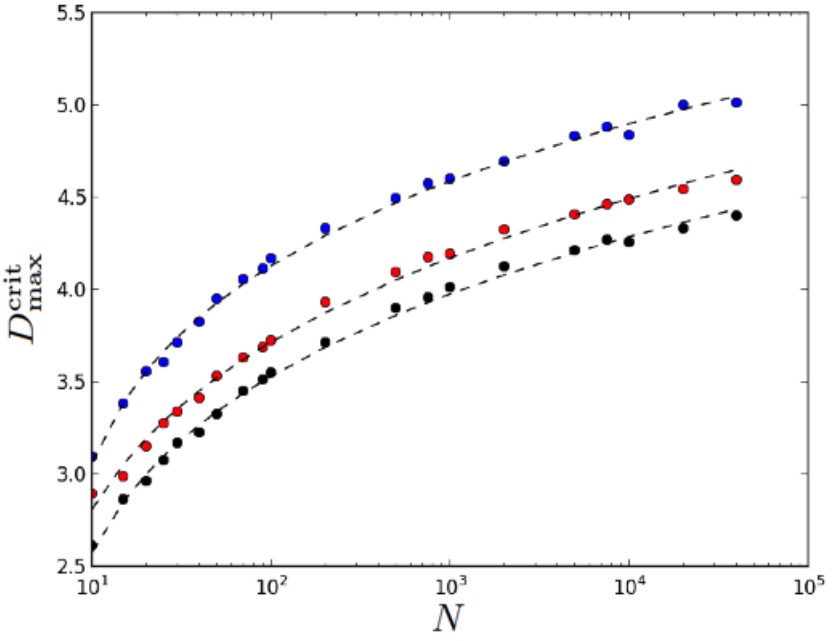


FIGURE 2.9: Critical curves for significance testing determined for a Gaussian distribution with the Anderson-Darling Test. Critical values of $D_{\max}$ as a function of the sequence length N, up to 40000, for series of numbers generated from a Gaussian distribution with $\sigma = 0.5$ and $\mu = 0$ performed with the Anderson-Darling test. The lines are fits to the data and the new parameters obtained for Eq. (2.3) are $(a, b, c) = (2.92, 1.47, 0.21)$.

We again perform tests on individual samples of Cauchy distributed segments of length N. A fit has been made and gives us new values for the parameters (a, b, c) for the significance that are now much larger due to the weighting. For $P_0 = 0.95$ we have

$$(a, b, c) = (2.92, 1.47, 0.21). \tag{2.7}$$

The curves obtained for Gaussian, Fig. 2.6, log-normal in Fig. 2.7, Cauchy in Fig. 2.8 with the KS algorithm and with the AD implementation for Cauchy distribution in Fig. 2.9 are more restrictive than the ones for the standard two-sample KS test of Fig. 2.2 that for large N tend to 1.22, 1.36 and 1.63 for confidence levels of 90%, 95% and 99%, respectively. The use of the classic KS test would lead to oversegmentation. This is due to the fact that in our case the two samples under comparison are not from random independent samples but arise from a cut within one set of data, hence the more restrictive criterion defined in Eq. (2.3) must be used and as shown works very well for the different classes of distributions tested. In the following section we implement the KS segmentation algorithm iteratively to a time series instead of only once.

2.5 Iteration

In the previous numerical tests the algorithm was only applied once to a data set. In this section it is applied iteratively to a generated time series. The interest of applying the KS segmentation algorithm recursively is that it runs through a whole time series the necessary number of times and fragments, when significant, until there is no more relevant cut position found. If we have a time series of length N the algorithm runs through it in the first iteration and finds one relevant cut position resulting in that the time series is now fragmented in two segments; in a second iteration the algorithm runs through both segments and again looks for significant cut positions and performs them if relevant. This goes on recursively until the size of one resulting fragment is smaller than the minimum size defined or if there are no more cut positions found.

2.5.1 Description of the Algorithm

Given an artificial time series of length N the algorithm visits the array $[0, N]$ and looks for a relevant position where it is significant to cut with the condition that $D_{KS} > D_{\max}^{\text{crit}}$ for the chosen significance and it takes the maximum of this value for all positions, and as a final requirement we impose that the length of the segment is greater than a predefined minimum length, l_0. The Significance is calculated with Eq. (2.3). Now suppose it finds significant cut position at position `*ptr1` pointed by the first red arrow. The dotted line indicates the position where the original time series is then fragmented resulting in two segments, `[0, *ptr1]` and `[*ptr1, N]`. At this point we have two time series and the KS algorithm runs through each of the segments. In the second iteration it has to visit the two fragments, `[0, *ptr1]` and `[*ptr1, N]` while the first visited array, `[0, N]`, is stored in the "visited sites". Next it finds a cut in the fragment `[0, *ptr1]` and none in the second array. Visited arrays are now `[0, N]`, `[0, *ptr1]` and `[*ptr1, N]`. After the cut position found at `*ptr2`, pointed by the second red arrow, the program now needs to process the two new resulting fragments `[0, *ptr2]` and `[*ptr2, *ptr1]`. The process ends when there is no cut position suggested in any of the resulting fragments or that one of the resulting fragment is smaller than the minimum length requirement, l_0. This procedure is illustrated in Fig 2.10 and the pseudo-code can be found in Appendix A. The code implementation was optimised by implementation of linked lists, which means that instead of the program sorting the whole fragments after each iteration it now takes the next pointer position and finds where it should be inserted in the previously sorted data array, this method greatly reduced execution speed, we do not know if a similar acceleration scheme was used in [3].

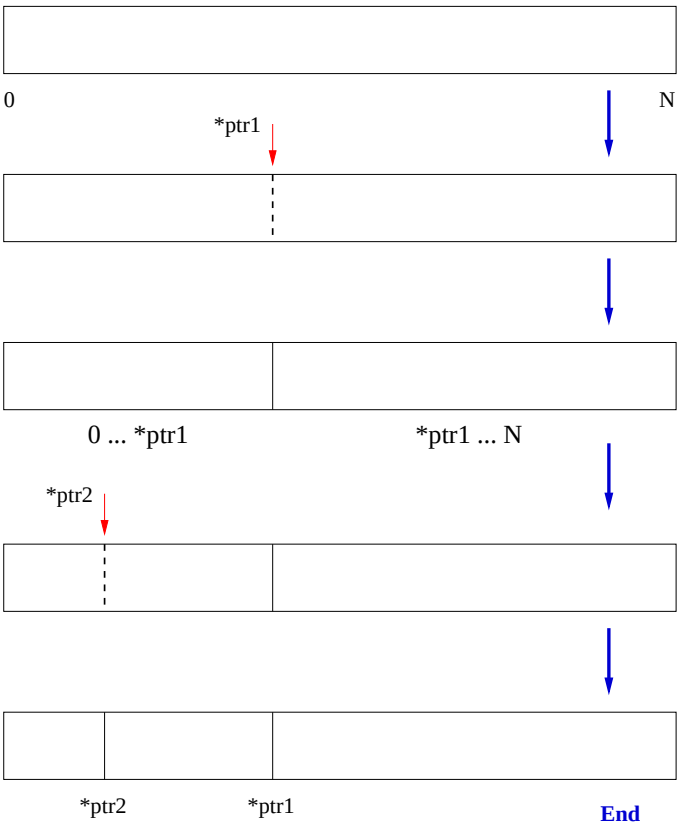


FIGURE 2.10: Illustration of iteration of the KS algorithm for a time series of length N. The diagram shows at first an unsegmented time series of size N, after a first iteration occurs a cut at position ***ptr1**, to where the first red arrow points, resulting in two segments of the initial time series. In the next step the algorithm runs on both segments and we show the example that it finds a relevant cut position in the left segment at ***ptr2**, to where the second red arrow points, and no cut on the right fragment. Then again it runs trough the resulting fragments and no more cuts are found relevant. This ends the procedure of the shown illustration for the iteration process. The algorithm goes on until there is no significant cut to be made on any of the resulting fragments or one of the resulting fragment is smaller than the minimum length requirement, l_0.

2.5.2 Numerical Tests for Segmentation Efficiency

To study the performance of the algorithm we analyse artificial time series formed by segments of m random numbers with alternating standard deviations σ_1 and σ_2 and alternating means with unitary jumps in consecutive segments. First we check the performance on a single time series generated from Gaussian random numbers with alternating means of $+0.5$ and -0.5 and standard deviation $\sigma_{1,2} = 0.2$ in Fig. 2.11 and different standard deviations $\sigma_1 = 1.0$ and $\sigma_2 = 3.0$ in Fig. 2.12. At this stage we have a visual idea of how accurate and efficient the KS segmentation algorithm is for certain values for standard deviations. This motivates us to investigate the performance of the algorithm for a large range of parameters and plot the results in the parameter plane which is done

in Subsection 2.5.2.2 for Gaussian, log-normal and Cauchy distributions with the KS and AD tests.

2.5.2.1 Testing Artificial Time Series

We generate an artificial time series of length 4000 with segments of 200 Gaussian numbers with alternating means of +0.5 and -0.5 and standard deviation $\sigma_{1,2} = 0.2$. Segmentation is performed at $P_0 = 0.95$ by means of the KS segmentation algorithm, for which we register for each iteration the cut position, i_{max}, the start and end positions of each resulting fragment and the respective length. Also we show the calculated values for D_{max}, $D_{\max}(i_{\max})$ and the significance calculated from Eq. (2.3) with the corresponding set of parameters $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$, and $(1.72, 1.86, 0.13)$ for the chosen significance level, for which we obtain the data shown in Table 2.2.

i_{max}	Start	End	Length	D_{max}	$D_{KS}(i_{\max})$	$D_{\max}^{\text{crit}}$
199	0	3999	4000	7.32	0.53	1.98
3800	199	3999	3801	6.88	0.5	1.97
3600	199	3800	3602	7.27	0.53	1.97
400	199	3600	3402	6.82	0.5	1.97
600	400	3600	3201	7.36	0.54	1.97
800	600	3600	3001	6.81	0.5	1.96
1000	800	3600	2801	7.4	0.54	1.96
1200	1000	3600	2601	6.76	0.5	1.96
1400	1200	3600	2401	7.47	0.55	1.95
1600	1400	3600	2201	6.77	0.5	1.95
1800	1600	3600	2001	7.61	0.57	1.94
3400	1800	3600	1801	6.62	0.5	1.94
3595	3400	3600	201	1.94	0.81	1.81
2000	1800	3400	1601	7.43	0.56	1.93
2200	2000	3400	1401	6.6	0.5	1.93
3200	2200	3400	1201	7.61	0.59	1.92
2400	2200	3200	1001	6.22	0.49	1.91
2601	2400	3200	801	8.1	0.66	1.9
2800	2601	3200	600	5.87	0.51	1.88
3000	2800	3200	401	9.86	0.99	1.86
3193	3000	3200	201	1.97	0.71	1.81

TABLE 2.2: Obtained results for the numerical testing on the efficiency of the segmentation algorithm. The KS segmentation algorithm is tested on an artificial time series of length 4000 with segments of 200 Gaussian numbers with alternating means of +0.5 and -0.5, standard deviation $\sigma_{1,2} = 0.2$ and segmentation is performed at $P_0 = 0.95$. For each iteration are registered the values of the cut position, i_{max}, the start and end positions of each resulting fragment and the respective length, the maximal distance, D_{max}, the cut acceptance criterion given by the critical value for the D_{max} value, $D_{KS}(i_{\max})$, and the cut acceptance criterion, $D_{\max}^{\text{crit}}$, calculated by means of Eq. (2.3). The number of proposed cuts is 21 while we expect 20, which shows that the KS segmentation algorithm is accurate with the parameters used. Moreover, D_{max} is almost always much larger than the $D_{\max}^{\text{crit}}$ needed for accepting a suggested cut at 5% significance level.

Looking at the Length column in Table 2.2 we see that the segments become shorter until the iteration terminates and if we compare columns $D_{\max}$ with $D_{\text{KS}^{\max}}$, $D_{\max}$ is always much larger than $D_{\text{KS}^{\max}}$. To illustrate the result of the KS segmentation algorithm on the tested time series we plot the time series in Fig. 2.11 and overlap the cut positions made by the algorithm, indicated in Table 2.2 in the column labeled i_{max} shown as vertical lines.

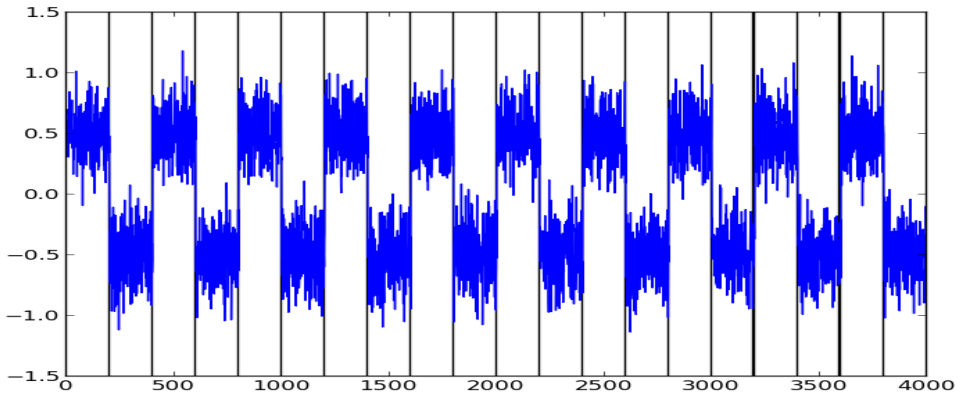


FIGURE 2.11: Illustration of the result of the segmentation algorithm on the tested artificial time series. The artificial time series formed by segments of $N = 200$ Gaussian numbers with alternating means +0.5 and -0.5 and standard deviation $\sigma_{1,2} = 0.2$. Segmentation was performed with the KS algorithm at $P_0 = 0.95$. The vertical lines indicate the output cut positions.

The algorithm cuts 21 times in all the positions it should plus one in a repeated position at 3200. It shows a good performance of the algorithm at detecting differences in the time series with the set of parameters used for $P_0 = 0.95$ significance level.

Next we choose different values for the standard deviation in consecutive segments. We generate an artificial time series with alternating means of +0.5 and -0.5 and standard deviation $\sigma_1 = 1.0$ and $\sigma_2 = 3.0$. While for the previous case we had the same standard deviation of $\sigma_{1,2} = 0.2$ in consecutive segments, as we can see in Fig. 2.11 the data does not disperse much from the mean values, making the limits between consecutive segments clear to the eye and to the algorithm. In this case we chose different standard deviations resulting in segments with $\sigma_2 = 3.0$ where the data values largely dispersed from the mean value making it less clear where each segment ends and another starts. We are interested in how the KS segmentation algorithm handles this choice of parameters.

The KS segmentation algorithm is performed again at $P_0 = 0.95$, for which we register for each iteration the cut position, i_{max}, the start and end positions of each resulting fragment and the respective length. Also the calculated values for D_{max}, $D_{KS}(i_{\max})$ and the cut acceptance criterion, $D^{\text{crit}}_{\max}$, by means of Eq. (2.3) with the corresponding set of parameters $(a, b, c) = (1.41, 1.74, 0.15)$, $(1.52, 1.8, 0.14)$ for the chosen significance level, for which we obtain the data shown in Table 2.3.

i_{max}	Start	End	Length	D_{max}	$D_{KS}(i_{\max})$	$D_{\max}^{\text{crit}}$
198	0	3999	4000	3.19	0.23	1.98
387	198	3999	3802	2.85	0.21	1.97
598	387	3999	3613	2.94	0.21	1.97
795	598	3999	3402	3.24	0.24	1.97
1003	795	3999	3205	3.01	0.22	1.97
1189	1003	3999	2997	2.43	0.18	1.96
1385	1189	3999	2811	2.73	0.2	1.96
1598	1385	3999	2615	2.92	0.21	1.96
1799	1598	3999	2402	3.21	0.24	1.95
3837	1799	3999	2201	2.26	0.18	1.95
1985	1799	3837	2039	2.42	0.19	1.95
2206	1985	3837	1853	2.64	0.19	1.94
2397	2206	3837	1632	2.6	0.2	1.93
2602	2397	3837	1441	3.27	0.25	1.93
2728	2602	3837	1236	2.87	0.27	1.92
3600	2728	3837	1110	2.34	0.17	1.92
3608	3600	3837	238	1.89	0.68	1.82
3405	2728	3600	873	2.87	0.23	1.9
3540	3405	3600	196	1.9	0.29	1.81
3200	2728	3405	678	2.43	0.2	1.89
3001	2728	3200	473	3.72	0.35	1.87
2773	2728	3001	274	2.4	0.39	1.83
517	387	598	212	1.88	0.27	1.82
391	387	517	131	1.86	0.94	1.78

TABLE 2.3: Obtained results for the numerical testing on the efficiency of the segmentation algorithm. The KS segmentation algorithm is tested on an artificial time series of length 4000 with segments of 200 Gaussian numbers with alternating means of +0.5 and -0.5, standard deviations $\sigma_1 = 1.0$, $\sigma_2 = 3.0$ and segmentation is performed at $P_0 = 0.95$. For each iteration are registered the values of the cut position, i_{max}, the start and end positions of each resulting fragment and the respective length, the maximal distance, D_{max}, the cut acceptance criterion given by the critical value for the D_{max} value, $D_{KS}(i_{\max})$, and the cut acceptance criterion, $D_{\max}^{\text{crit}}$. The number of proposed cuts is 24 while we expect 20, which shows that the KS segmentation algorithm is not as accurate as the previous case where is used standard deviations $\sigma_{1,2} = 0.2$.

Looking at the Length column in Table 2.2 we see that the segments become shorter until the iteration terminates and if we compare columns $D_{\max}$ with $D_{\text{KS}^{\max}}$, $D_{\max}$ is almost always much larger than $D_{\max}^{\text{crit}}$. In Fig. 2.12 is plotted the artificial time series generated overlapped with the cut positions, indicated in Table 2.2 as i_{max}, where we can see an illustration of the segmentation algorithm result.

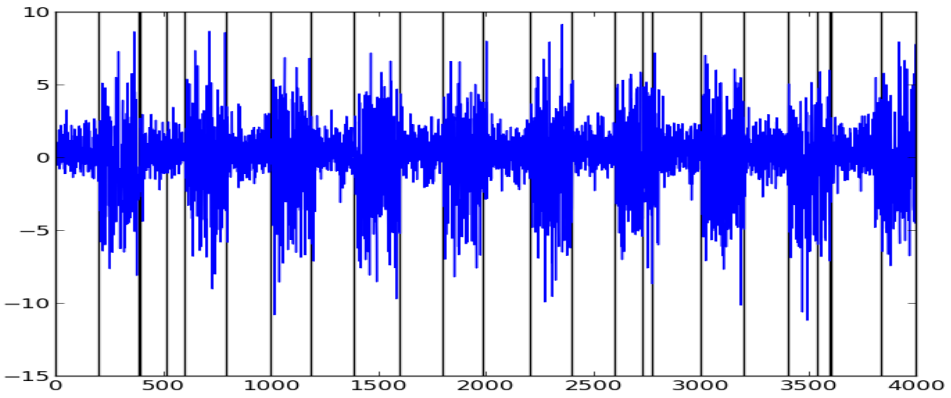


FIGURE 2.12: Illustration of the result of the segmentation algorithm on the tested artificial time series. The artificial time series if formed by segments of $N = 200$ Gaussian numbers with alternating means +0.5 and -0.5 and standard deviation $\sigma_1 = 1.0$ and $\sigma_2 = 3.0$. Segmentation was performed with the KS algorithm at $P_0 = 95\%$. The vertical lines indicate the output cut positions.

Fig. 2.12 shows us that the choice made in standard deviation values, $\sigma_1 = 1.0$ and $\sigma_2 = 3.0$, lead to oversegmentation of the algorithm due to the fact 24 cuts are proposed while only 20 should exist. Now that we have a picture of the efficiency of the KS segmentation algorithm we want to make a deeper investigation into the range of parameters where the algorithm performs better and worse which we do by plotting diagrams of the segmentation results in the parameter plane. In Fig. 2.11 and Fig. 2.12 we have reproduced the analysis found in [3].

2.5.2.2 Testing the Performance of the Algorithm

We are interested in evaluating the performance efficiency of the KS segmentation algorithm which is done analysing artificial time series formed by segments of m random numbers of Gaussian, log-normal and Cauchy distributions. Diagrams of the segmentation results in the parameter plane σ_1, σ_2 are shown for different distributions and parameters, and for each sequence, the relative number of cuts with respect to the actual one is displayed in color scale. We also introduced the minimum length requirement for a given segment, defined as l_0, following the approach presented in [3], to be evaluated before the algorithm makes a cut. Time series belonging to the middle of the color scale, the green area, are correctly segmented (100%), while those part of the blue area are typically unsegmented (0%) and the orange to red regions are oversegmented (>100%). The metric is the result of the number of cuts performed by the algorithm divided by the correct number of cuts. For example if the algorithm makes 24 cuts while the correct

number of cuts is 20 our metric is given by $24/20 \times 100\% = 120\%$ indicating oversegmentation and shown in the orange-red area of the parameter plane. Diagrams of the KS segmentation algorithm results in the plane σ_1, σ_2 are shown in Figs. 2.13, 2.16 and 2.17 for $l_0 = 10$, $l_0 = 50$ and $l_0 = 0$, respectively, for Gaussian time series of random numbers with segment size of $m = 200$. In Fig. 2.18 is shown the diagram for the segmentation algorithm in the parameter plane for log-normal distribution with and in Fig. 2.21 for the case of random numbers generated from a Cauchy distribution both for $l_0 = 10$ and $m = 200$. With these results we compare the performance of the KS segmentation algorithm between its application to time series composed of random numbers from a Gaussian and a Cauchy distribution in Fig. 2.23 and see that it works better for the Gaussian case. Next we will make the comparison between the KS segmentation algorithm and its modified version, the AD implementation which adds a weight parameter in order to maximize the sensitivity at the tails of the distribution. In fact it will be seen in Fig. 2.25 that the AD performs slightly better than the KS version for the Gaussian distribution case. In all cases, each grid cell corresponds to a different random sequence of size $N = 4000$ and segment size m. Segmentation is performed with the KS algorithm with a new condition before final acceptance of the cut, we require a minimal size of a fragment, l_0, namely $i_{\max} - i_1 + 1, i_n - i_{\max} \geq l_0$.

2.5.2.3 Performance of the Algorithm for Gaussian distribution

We set the segment size at $m = 200$ and minimal length requirement $l_0 = 10$ and test for different confidence level values P_0 at $0.90, 0.95$ and 0.99 for Gaussian random numbers with mean values one unit apart, and obtain for each the diagrams of Fig. 2.13.

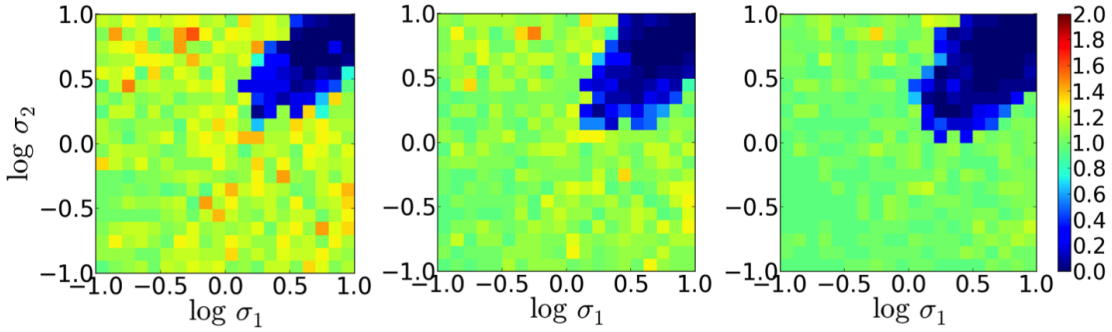


FIGURE 2.13: Performance of the KS segmentation algorithm for Gaussian distribution at different significance levels. Segmentation diagram in the parameter plane σ_1, σ_2 on log-log axes. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with minimum length requirement of $l_0 = 10$ and $P_0 = 0.90$ (on the left), $P_0 = 0.95$ (in the middle) and $P_0 = 0.99$ (on the right).

It is clear from Fig. 2.13, the higher the significance level the better performs the algorithm. Light green cells correspond to 100% accurate segmentation while above, in the color palette, between yellow and red corresponds to over 100%, hence oversegmentation. The blue area, where the algorithm is under 100% accuracy, under segmentation, gets slightly larger the higher the significance level chosen. Even then, the best performance here is seen on the right diagram of Fig. 2.13 at $P_0 = 0.99$ where the number of wrong segmentation cases is lower.

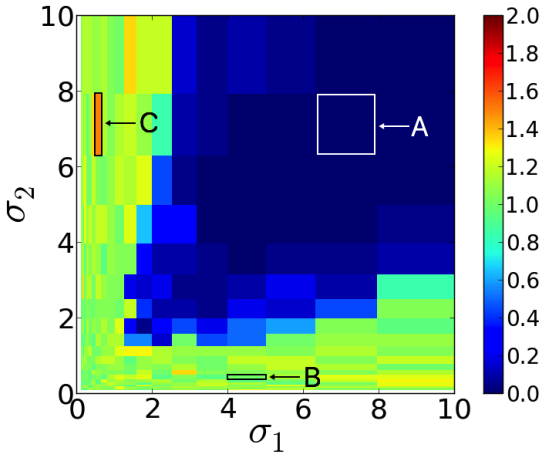


FIGURE 2.14: Performance of the KS segmentation algorithm for Gaussian distribution at $P_0 = 0.95$. Segmentation diagram in the parameter plane σ_1, σ_2. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 10$ and $P_0 = 0.95$, corresponding to the center diagram of Fig. 2.13, but now in a linear scale.

In Fig. 2.14 is shown the parameter plane of the standard deviations, σ_1, σ_2. We choose from this plot three cells, each of which corresponds to a different case, a blue cell shows undersegmentation, green is correct segmentation and red shows oversegmentation. In Fig. 2.15 we show time series and distributions for sets of parameters that result in time series that are unsegmented, correctly segmented and oversegmented corresponding to blue, green and red cells, respectively, in the parameter plane shown in Fig. 2.14.

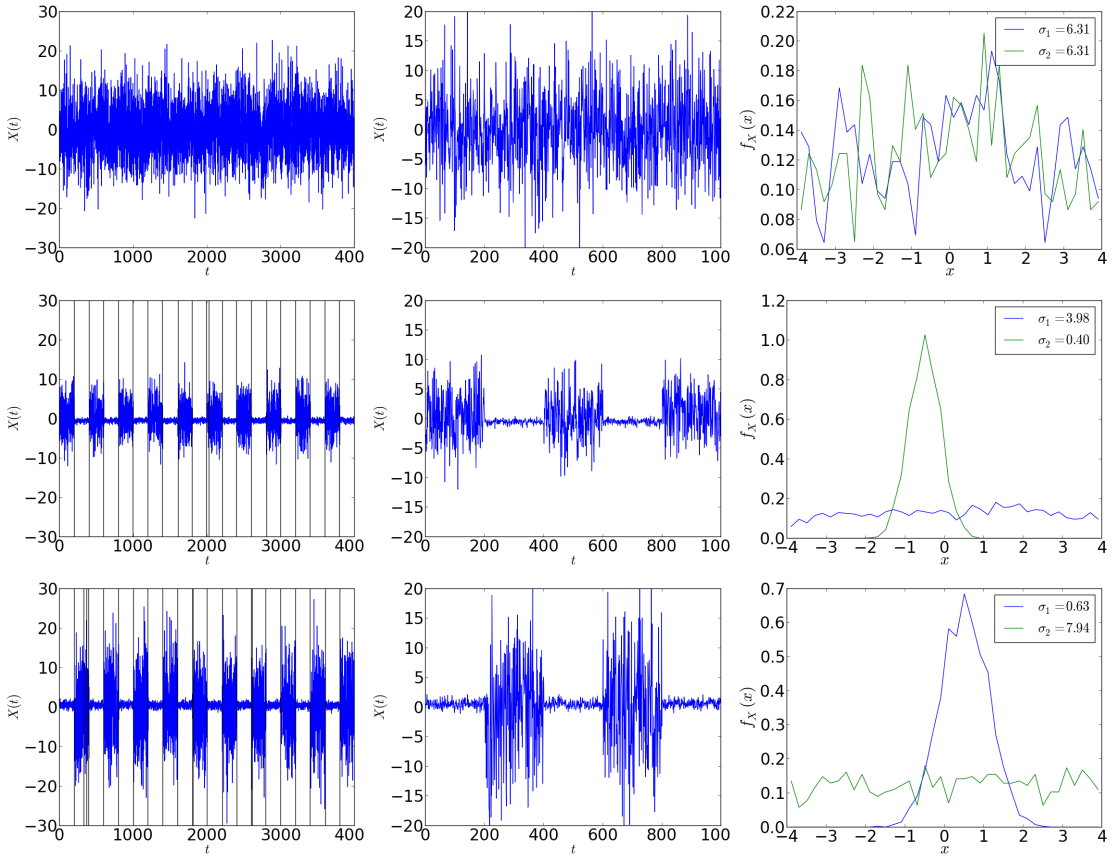


FIGURE 2.15: For three points, A, B, C, in the parameter plane of Fig. 2.14, corresponding to the performance of the KS segmentation algorithm for a Gaussian distribution, we show time series and distributions for sets of parameters that result in time series that are unsegmentable, shown in the blue region by cell A, correctly segmented (100%) for the green cell marked by B and oversementated (>100%) marked by cell C. For each of these we show in the left the complete composite artificial time series with the cuts performed by the KS segmentation algorithm shown as vertical black lines, in the middle is the same time series but zoomed in and on the right are shown the empirical PDFs for the two distributions with corresponding standard deviations in legend. The first row corresponds to the blue cell indicated by point A ($\sigma_1 = 6.31, \sigma_2 = 6.31$) where the left figure shows zero cuts made by the segmentation algorithm while the middle plot shows a fragment of the time series where is clear that it is difficult to see a difference between segments which is illustrated in the right plot where we see strongly overlapping PDFs. On the second row is the case of correct segmentation corresponding to the green cell indicated by B in Fig. 2.14 at $\sigma_1 = 3.98$ and $\sigma_2 = 0.40$. The left figure shows correct segmentation of the time series aside from one additional one around 2000. The middle figure shows the partial time series where now consecutive segments are distinguishable. In the right figure we can see that the PDFs are not overlapping as much as the previous case. The third row, shown as the red cell C in Fig. 2.14, indicates oversegmentation illustrated by a red cell with parameters $\sigma_1 = 0.63, \sigma_2 = 7.94$. In the left figure we see that the time series is segmented where it should but has many additional cuts. The middle figure shows a fragment of the time series and the right shows us slightly overlapping PDFs.

Each row Fig. 2.15 refer to a pair of standard deviation values σ_1 and σ_2 indicated by A, B, C in the parameter plane Fig. 2.14, for each we show in the first column the complete time series with the cuts performed by the KS segmentation algorithm illustrated with

vertical black lines, in the second column is plotted the same tested time series but zoomed to have a visual idea of the composite time series and in the third row are the PDFs for each of the standard deviation values σ_1 and σ_2.

We want to explore what is the influence on the accuracy of the algorithm if the minimum length requirement is changed.Therefore we increase the minimum length requirement to $l_0 = 50$ and test for different confidence level values P_0 at $0.90, 0.95$ and 0.99 for Gaussian random numbers with segment size of $m = 200$ and obtain for each confidence level the diagrams of Fig. 2.16.

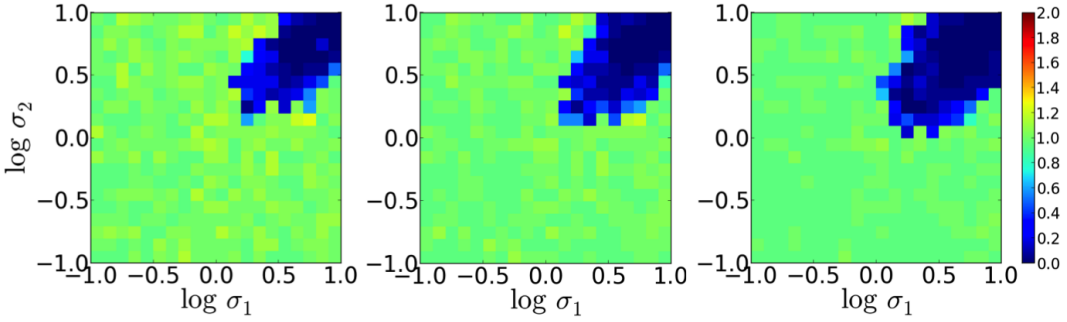


FIGURE 2.16: Performance of the KS segmentation algorithm for Gaussian distribution at different significance levels. Segmentation diagram in the parameter plane σ_1, σ_2 on logarithmic scale. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 50$ and $P_0 = 0.90$ (on the left), $P_0 = 0.95$ (in the middle) and $P_0 = 0.99$ (on the right).

In comparison to Fig. 2.13 this case is considerably better given the absence of yellow, orange and red cells indicatives of oversegmentation. When we set a larger value for the minimum length requirement l_0 before a cut, small segments are discarded hence the number of false cuts is reduced corroborated by comparing Fig. 2.13 and Fig. 2.16. To validate this we set no minimum length, i.e, $l_0 = 0$ in Fig. 2.17 with the same test parameters of random sequences of size $N = 4000$ and segment sizes of $m = 200$, mean values one unit apart and for a confidence level of $P_0 = 0.95$ in Fig. 2.17 to compare with the cases of $l_0 = 10$ in Fig. 2.13 and $l_0 = 50$ in Fig. 2.16.

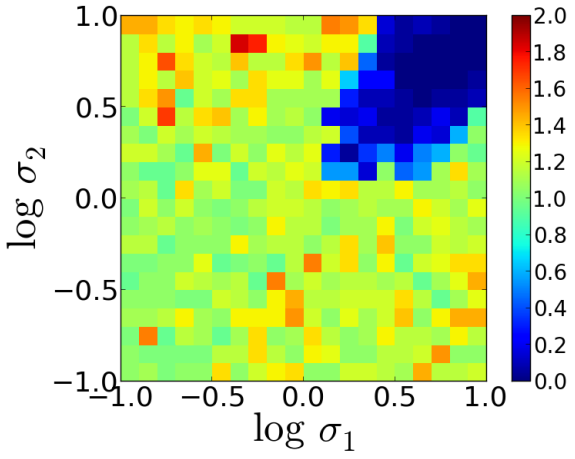


FIGURE 2.17: Performance of the KS segmentation algorithm for Gaussian distribution at at confidence level of $P_0 = 0.95$. Segmentation diagram in the parameter plane σ_1, σ_2 on logarithmic scale. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 0$ and $P_0 = 0.95$.

Comparing Fig. 2.17 where no minimum length l_0 is set with Fig. 2.13 and Fig. 2.16 where the minimum length requirement is $l_0 = 10$ and $l_0 = 50$, respectively, it is evident that this length requirement has an important role at reducing significantly the number of wrong cuts, avoiding mostly oversegmentation of time series corresponding to a ratio of number of performed cuts with the number of correct cuts greater than 100% shown by orange and red cells in the diagrams. In Fig. 2.17 we observe a much higher occurrence of orange to red cells which show oversegmentation higher than 140%. The worst performance of the algorithm occurs for the pair $\sigma_1 = 0.40, \sigma_2 = 6.31$ with 37 cuts leading to an oversegmentation of 185%. This case shows that without a minimum length requirement before cutting the KS segmentation algorithm performs very negatively presenting abundance of under and oversegmentation illustrated as blue and orange to red cells, respectively. In the next case we shift our attention to time series drawn from the log-normal distribution to initiate our performace testing of the KS segmentation algorithm on distributions which present different tail behaviour.

2.5.2.4 Performance of the Algorithm for log-normal distribution

A log-normal distribution is a continuous probability distribution of a random variable whose logarithm is normally distributed. If the random variable X is log-normally distributed then $Y = \ln(X)$ has a normal distribution. The PDF is given by Eq. (1.2.14). For small σ_Y the distribution is approximated by the normal distribution [24]. We set the segment size at $m = 200$ and minimal length requirement $l_0 = 10$ and test for different confidence level $P_0 = 0.95$ for log-normal random numbers and obtain the diagram

shown in Fig. 2.18. The location parameter values are chosen such that they are one unit apart, namely $\mu_{Y_1} = 1.0$ and $\mu_{Y_2} = 2.0$.

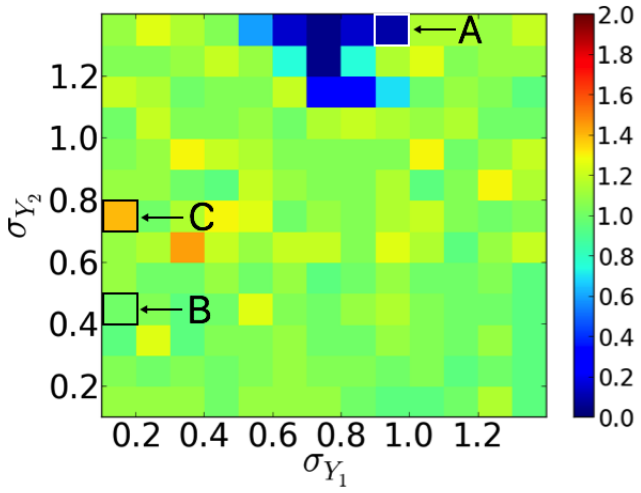


FIGURE 2.18: Performance of the KS segmentation algorithm for log-normal distribution. Segmentation diagram in the parameter plane $\sigma_{Y_1}, \sigma_{Y_2}$ in linear scale. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 10$ and confidence level $P_0 = 0.95$.

The algorithm seems to perform very well for a log-normal distribution given the lack of evidence of oversegmentation and undersegmentation illustrated by red and orange cells and blue region, respectively. For most pairs of $\sigma_{Y_1}, \sigma_{Y_2}$ the cells are green which result of correct segmentation of the corresponding time series, we observe very few blue cells for undersegmentation and even less orange cells indicating oversegmentation occurs rarely for the chosen range of parameters.

Note that $\sigma_{Y_1}, \sigma_{Y_2}$ in this case do not correspond to the standard deviations but to the scale parameter of the distribution. The standard deviation for the log-normally distributed random variable X is the variance squared given by Eq. (1.7):

$$\sigma_X = \sqrt{\mu_X^2 \left(e^{\sigma_Y^2} - 1 \right)} \tag{2.8}$$

The plot in Fig. 2.19 is the same result as in Fig. 2.18 but in the standard deviation plane $\sigma_{X_1}, \sigma_{X_2}$ calculated with Eq. (2.8), in order to facilitate comparison with the Gaussian case.

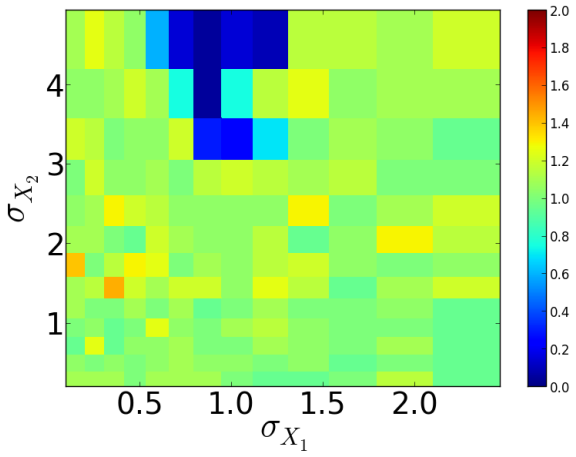


FIGURE 2.19: Performance of the KS segmentation algorithm for log-normal distribution. Segmentation diagram in the standard deviation plane for linear scale. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 10$ and $P_0 = 0.95$. The data is the same than in Fig. 2.18 but plotted on a different parameter scale.

In Fig. 2.20 we show time series and distributions for sets of parameters that result in time series that are unsegmentable ($< 100\%$), correctly segmented (100%) and over segmented ($> 100\%$) corresponding to blue, green and red cells, respectively, in the parameter plane shown in Fig. 2.18.

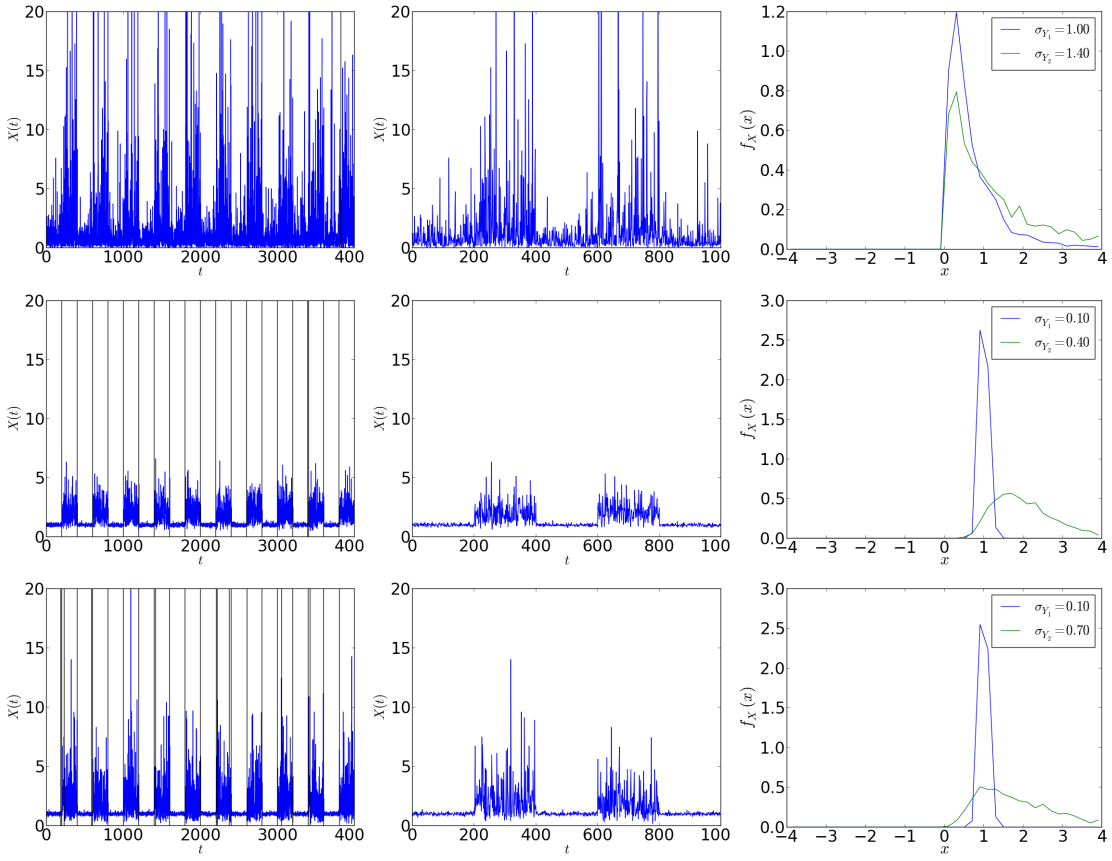


FIGURE 2.20: The three rows correspond to three points, A, B, C, in the parameter plane of Fig. 2.18, corresponding to the performance of the KS segmentation algorithm for a log-normal distribution, we show time series and distributions for sets of parameters that result in time series that are undersegmented, shown in the blue region (<100%) by point A, correctly segmented (100%) for the green cell indicated by B and oversementated (>100%) marked by orange cell C. For each row the left column displays the complete composite artificial time series with the cuts performed by the KS segmentation algorithm shown as vertical black lines, in the center column is the same time series but zoomed in and on the right column are shown the PDFs for the two distributions with corresponding standard deviations in legend. The first row corresponds to a blue cell in Fig. 2.18 at position $\sigma_1 = 1.00$ and $\sigma_1 = 1.40$. From the left figure we see that only one cut is performed approximately at position 3800 resulting in a performance of the algorithm of only 5%. Both standard deviations are in the upper limit of the range resulting in the time series shown in the middle figure where the different segments within the time series are not clear. The corresponding PDFs display similar shape and overlap. In the second row is presented the case of correct segmentation with $\sigma_1 = 0.10$ and $\sigma_1 = 0.40$ performing with 100% accuracy, i.e. 20 cuts, as can be seen in the left figure where the vertical black lines are positioned exactly where a new segment with different parameters starts. In the middle figure we see the zoomed time series and last the PDF which present different shapes and do not overlap as the above case resulting in correct segmentation. In the third row is shown the case with $\sigma_1 = 0.10$ and $\sigma_1 = 0.70$ illustrating the orange cell indicated by C in Fig. 2.18 corresponding to oversegmentation with 140% performance, this is obvious if we look at the left figure where there are a large number of repeated cut positions shown by overlapping vertical black lines. In the middle figure is shown the corresponding zoomed time series and in the right one are presented the PDFs.

Interpreting Fig. 2.18 and choosing three cells corresponding to sets of parameters σ_1, σ_2

leading to undersegmentation, A, correct segmentation, B, and oversegmentation, C, we observe that for higher values of σ_1, σ_2 the KS algorithm performs worse. This could be because higher σ_1, σ_2 values lead to heavier tails in a log-normal distribution for which the KS test is less sensitive. In Fig. 2.20 the third row has parameters $\sigma_1 = 0.10$ and $\sigma_1 = 0.70$ and performs 28 times, which shows a ratio of performed cuts/correct cuts of 1.40, leading to 140% performance. While for $\sigma_1 = 1.00$ and $\sigma_1 = 1.40$ the algorithm finds only 1 cut, instead of the 20 expected, as shown in the first figure of the first row in Fig. 2.20 leading to an accuracy of only 5%. In average the algorithm makes 20,14 cuts while the worst case is for $\sigma_1 = 0.30$ and $\sigma_1 = 0.60$ where it cuts 29 times resulting in 145% performance. Nevertheless, on the whole, the algorithm performs well for log-normally distributed time series.

We tested the KS segmentation algorithm on log-normally distributed samples and observed it performed good, now the next step is to test the algorithm on a distribution with heavier tails, the Cauchy distribution, described in 1.2.2.5.

2.5.2.5 Performance of the Algorithm for Cauchy distribution

Exploring how the test performs on data from a Cauchy distribution is a good indicator of its sensitivity to a heavy-tailed distribution, since the Cauchy distribution is unimodal like the Gaussian, but with a power law tail. Following the same procedure as before we compute the performance of the algorithm for the Cauchy distribution with PDF given by Eq. (1.10) with location parameter $x_0 = 0$ and range over the scale parameter γ. We generate 20 segments each with size $m = 200$ and total sequence size $N = 4000$ with the algorithm performing at $P = 0.95$ and minimum length requirement $l_0 = 10$. The range for the scale parameter is $\gamma_{min} = 0.1$ to $\gamma_{max} = 2.1$ with separations of 0.1.

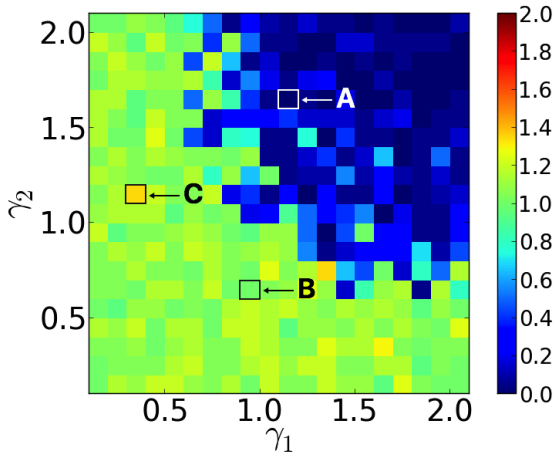


FIGURE 2.21: Performance of the KS segmentation algorithm for Cauchy distribution at a given significance level. Segmentation diagram in the parameter plane γ_1, γ_2 for linear scale. The relative number of cuts is represented in a color scale mapping. Each cell corresponds to a random sequence of size $N = 4000$ and segment sizes of $m = 200$. Segmentation is performed with $l_0 = 10$ and $P_0 = 0.95$.

In Fig. 2.21 the performance of the algorithm for Cauchy distribution is clearly defined. The blue zone is larger, however the plot is in the plane of the scale parameters. Under-segmentation would occur more frequently for scale parameter range $\gamma_1, \gamma_2 > 1$. Nevertheless the KS segmentation algorithm is capable of correctly segmenting time series from a Cauchy distribution in a large region of the parameter space.

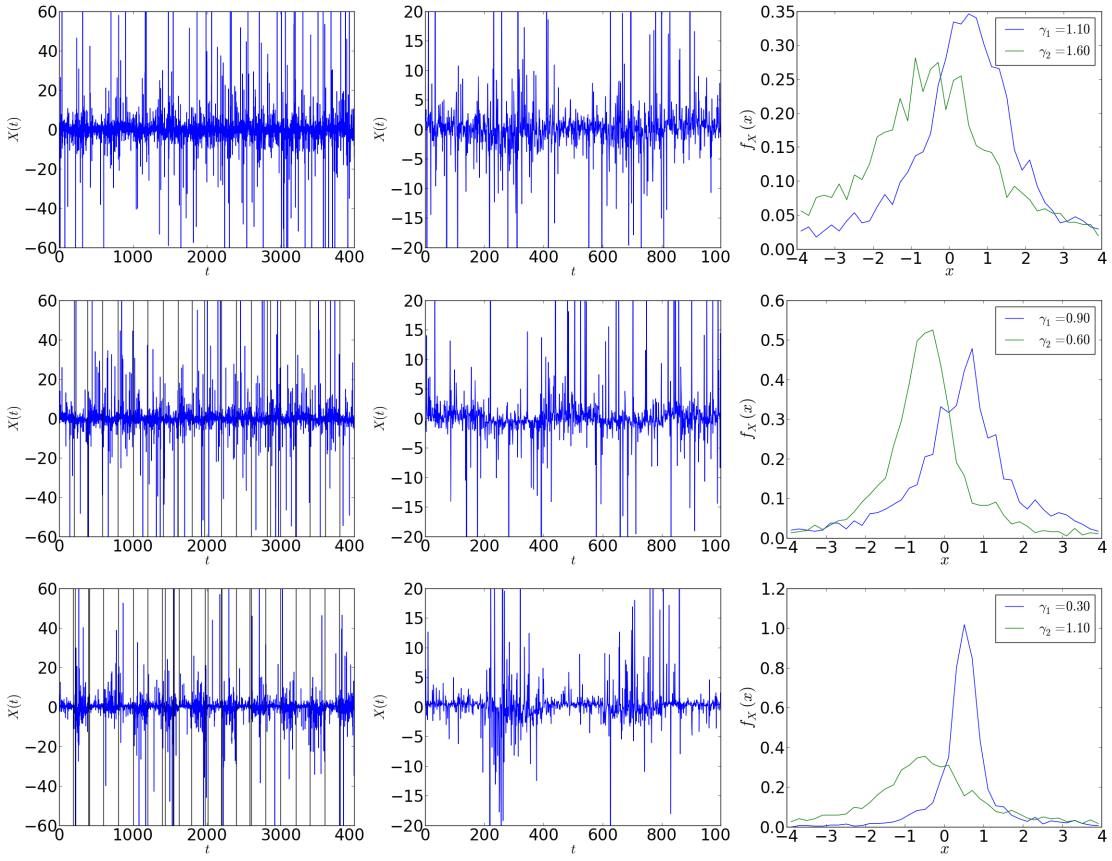
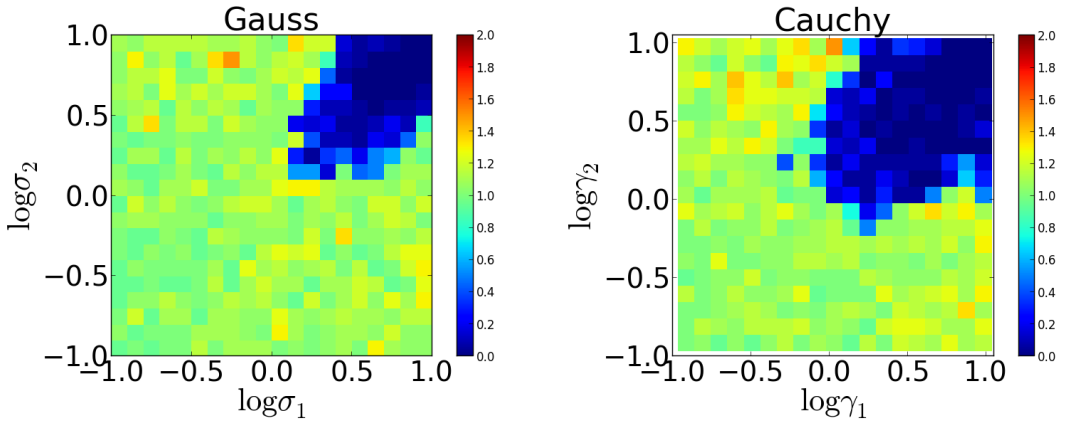


FIGURE 2.22: The three rows correspond to three points labeled by A, B, C in Fig. 2.21, corresponding to the performance of the KS segmentation algorithm for a Cauchy distribution, we show time series and distributions for sets of scale parameters (γ_1, γ_2) that result in time series that are undersegmented, A, shown in the blue region (<100%), correctly segmented (100%) for a green cell indicated by B and oversementated (>100%) marked by C. For each row the left column displays the complete composite artificial time series with the cuts performed by the KS segmentation algorithm shown as vertical black lines, in the middle column is the same time series but zoomed in and in the right column are shown the PDFs for the two distributions with corresponding scale parameters $\gamma_{1,2}$ in legend. The first row is relative to a cell in the dark blue region, cell A, with $\gamma_1 = 1.10$ and $\gamma_1 = 1.60$ where occurs undersegmentation, in fact for this set of parameters no cut is made which can be confirmed observing the left figure where no vertical black lines exist, hence the algorithm has a performance of 0%. The middle figure shows a amplification of the tested time series from where we cannot say where the segments start and end making the algorithms task of finding differences difficult. The right figure illustrates the corresponding PDFs which show overlap and noticeable tails. The second row exhibits time series and PDF for cell B where the algorithm performs with 100% accuracy cutting 20 times. This corresponds to a green cell with scale parameters $\gamma_1 = 0.90$ and $\gamma_1 = 0.60$. In the left figure is shown the complete time series with the 20 vertical cuts found and in the center figure is a close up on the time series. The right figure shows the corresponding PDFs which display similar shape. The third row illustrates an orange cell at $\gamma_1 = 0.30$ and $\gamma_1 = 1.10$, point C, which results in oversegmentation. This is clear by observing the left figure where in some positions are overlapping vertical black lines. The algorithm makes 27 cuts resulting in 135% cuts compared to expected ones. The middle figure shows the zoomed time series and the third the PDFs where the PDF for $\gamma_1 = 0.30$ has a much higher peak than the PDF for $\gamma_2 = 1.10$ and the right hand side tails of both distributions strongly overlap.

Interpreting Fig. 2.18 and choosing three cells corresponding to sets of parameters γ_1, γ_2 leading to undersegmentation, A, correct segmentation, B, and oversegmentation, C, we can see that for larger values of γ_1, γ_2 the KS algorithm performs badly. This could be because higher scale parameter values lead to heavier tails in a Cauchy distribution for which the KS test is less sensitive. In average the algorithm makes 14 cuts which is very low due to the fact that more than 50% of the cells belong to the blue region of undersegmentation cutting less than 20 times and from these 50%, 22% of the time series are cut less than 10 times rather than the correct 20. Correctly segmented time series (20 cuts) occur exactly 36 times representing only 8% of the total trials. It should be noted, however, that the parameter range here was chosen arbitrarily and no comparison with the Gaussian or log-normal case can be drawn without defining a common parameter describing the distribution width. This is a complicated challenge, since the Cauchy distribution possesses no finite second moment.

2.5.2.6 Comparison of the Performance between Gaussian and Cauchy cases

Since we want to compare different distribution classes with completely different tail behaviour, such as Gauss and Cauchy, we need a comparable measure of the width of the distributions. Because of the fact that Cauchy distribution does not have defined mean or standard deviation we need another measure for the PDF width. We use the $1/e$ width, which is the value of the argument of the PDF, measured from the maximum, for which the PDF has decayed to $1/e$ of its maximum, plotted in Fig. 2.24. In Fig. 2.23 is shown the comparison of the performance of the KS segmentation algorithm for Gaussian and Cauchy distributions with the same parameter range for σ_1, σ_2 and γ_1, γ_2.



(A) (B)

FIGURE 2.23: Comparison of the performance of the KS segmentation algorithm between Gaussian and Cauchy distributions for $P_0 = 0.95$ confidence level and minimum length requirement $l_0 = 10$. Figure (A) shows the segmentation diagram in the parameter plane for the Gaussian case with axes corresponding to the logarithms of the standard deviations; On the right hand side, figure (B), is the segmentation diagram in the parameter plane for the Cauchy distribution with the axes corresponding to the logarithm of the scale parameter. For the same range of the parameters σ and γ the unsegmented blue area is larger for the Cauchy case, figure (B).

We want to write the PDF for a Gaussian distribution in terms of the $1/e$ measure and taking into consideration the distribution has its maximum at $x = 0$

$$\frac{1}{e} = e^{-\frac{x^2}{2\sigma^2}} \tag{2.9}$$

solving for x we obtain

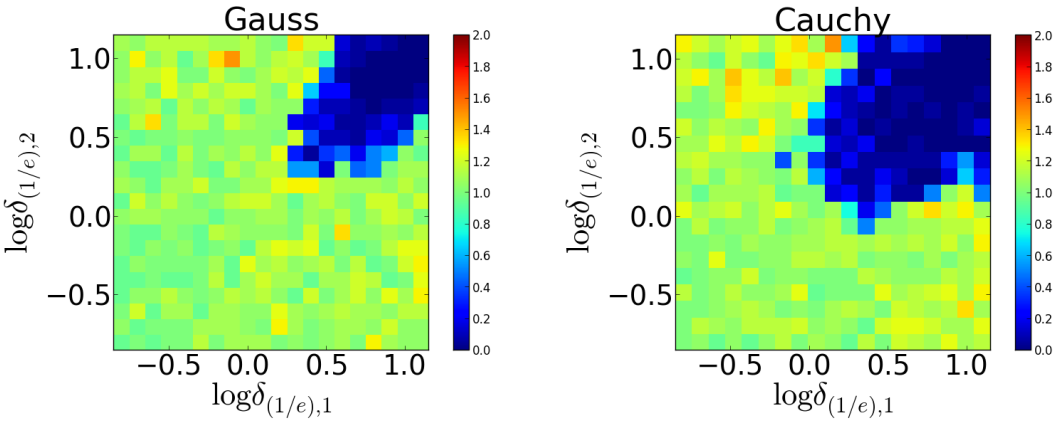
$$x = \sqrt{2}\sigma = \delta_{(1/e)}. \tag{2.10}$$

And the PDF for a Cauchy distribution in terms of the $1/e$ measure

$$\frac{1}{e} = \frac{1}{\gamma\pi\left[1 + \frac{x^2}{\gamma^2}\right]} \tag{2.11}$$

solving Eq. (2.11) for x and taking into consideration the distribution has its maximum at $x = \frac{1}{\gamma\pi}$ we get that

$$x = \gamma\sqrt{e-1} = \delta_{(1/e)}. \tag{2.12}$$



(A) (B)

FIGURE 2.24: Comparison of the performance of the KS segmentation algorithm between Gaussian and Cauchy distributions for $P_0 = 0.95$ confidence level and minimum length requirement $l_0 = 10$ in terms of the $1/e$ length, $\delta_{(1/e)}$. Figure (A) shows the segmentation diagram in the parameter plane for the Gaussian case with axes corresponding to the logarithms of the $1/e$ length, $\delta_{(1/e)}$; On the right hand side, figure (B), is the segmentation diagram in the parameter plane for the Cauchy distribution with the axes corresponding to the logarithm of the corresponding $1/e$ length, $\delta_{(1/e)}$.

From Fig. 2.24 we observe that in terms of the $1/e$ length, $\delta_{(1/e)}$, for the same range of parameter values, for Cauchy distributed time series, Fig. 2.24 (B), the blue area is larger indicating more undersegmentation than for Gaussian distributed time series, Fig. 2.24 (A). Moreover, for the Gaussian case we see more cells belonging to the yellow area of the color scale of Fig. 2.24 which indicate slight oversegmentation of approximately 20%. Note that we cannot make judgements about the performance of the algorithm comparing Gaussian and Cauchy cases because even if the parameter range is the same, the parameters themselves are different and individual to their respective distribution and the distributions itself are consequently different, however this comparison gives us the idea of how the algorithm behaves for the same range of parameters for both distributions. Table 2.4 summarises the obtained results.

Analysing Table 2.4 we see that for Cauchy distributions we obtained less oversegmentation but more undersegmentation while for Gaussian distributions, the contrary, more oversegmentation and less undersegmentation. For Cauchy distributions we obtained more correct cuts than for the Gaussian case which is surprising, meaning the algorithm works well even for a distribution with tail behaviour. This affirms our motivation to seek improvement in the sensitivity of the algorithm near the tails of a distribution for better performance.

TABLE 2.4: Summary of results for the comparison of the performnce of the KS segmentation algorithm between Gaussian and Cauchy distributions. The table shows the number of occurences and the matching percentage that correspond to oversegmentation with more than 20 and more than 25 cuts, indicating extreme oversegmentation, undersegmentation with under 20 and under 10 cuts, indicating extreme undersegmentation, and the number of correct segmentation with exactly 20 cuts. The two last rows show extreme situations of over, > 25, and undersegmentation, < 10. The table translates the main features observed in Fig. 2.24, essentially that for Cauchy distributions we have less oversegmentation than for Gaussian distributions and for Cauchy class of distributions the algorithm cuts correctly more times than for Gaussians, namely 10.0% against 5.0%.

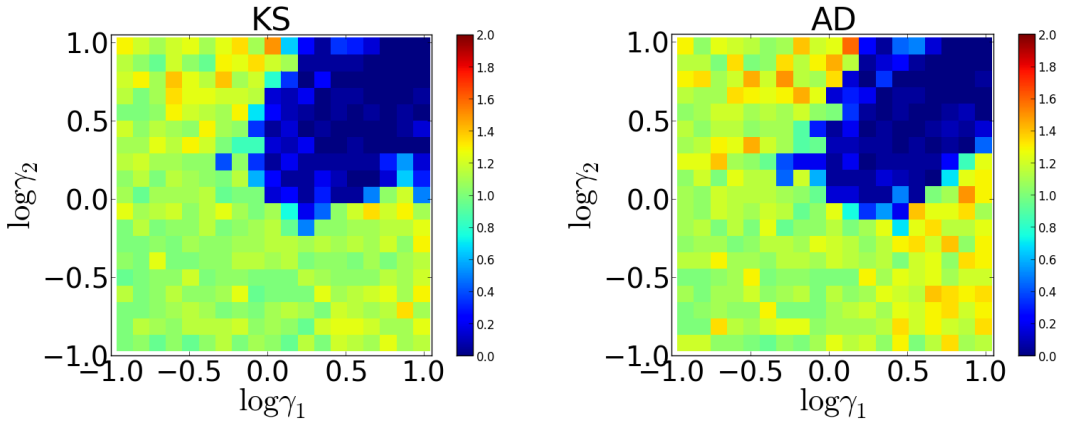
	Gauss		**Cauchy**	
	#	%	#	%
$>$20	327	74.1	228	57.0
$<$20	92	20.9	132	33.0
$=$20	22	5.0	40	10.0
$>$25	93	21.1	22	5.5
$<$10	64	14.5	108	27

2.5.2.7 Comparison of the Performance between KS and AD for Cauchy distribution

In Fig. 2.23 we compared the result of the KS segmentation algorithm on time series from Gaussian and Cauchy distributions and observed that the segmentation algorithm based on the KS test performs with positive results on both cases. Cauchy distributions are unimodal like Gaussian distributions, see Fig. 1.1 and Fig. 1.3 for PDFs of Gaussian and Cauchy distributions, respectively. Exploring how a test performs on time series from a Cauchy distribution is a good indicator of how sensitive the test is to heavy tail departures from normality. This prompts us to apply the AD test according to Subsection 1.6.1 introducing a weight term which translates into a higher sensitivity to tail behaviour. We compare the performances of the KS and AD tests for time series belonging to Cauchy distributions in terms of the $1/e$ length with a confidence level of $P_0 = 0.95$. The cut acceptance criterion is calculated for both cases with

$$D_{\max}^{\text{crit}}(N_e) = a(\ln N_e - b)^c.$$

For the AD test the parameters (a, b, c) used are the ones obtained when testing the significance criterion for the AD test in Subsection 2.4.2.4, for $P_0 = 0.95$ we have $(a, b, c) = (2.92, 1.47, 0.21)$.



(A) (B)

FIGURE 2.25: Comparison of the performance of the KS segmentation algorithm and the modified AD algorithm applied to time series belonging to Cauchy distributions. In (A) is the segmentation diagram in the parameter plane for the Cauchy case with the axes corresponding to the logarithm of the scale parameter in terms of the $1/e$ length, performed with KS test and in (B) is implemented the AD modification. The modified AD test performs slightly better than the KS segmentation algorithm.

Fig. 2.25 illustrates the performance of the segmentation algorithm based on the KS, Fig. 2.25 (A), and AD, Fig. 2.25 (B), tests. Comparing the two figures we see a slight but not too significant increase in performance. The blue area in Fig. 2.25 (B) is smaller to (A) indicating that undersegmentation would occur less frequently, however there exist more orange cells which correspond to slight oversegmentation of approximately 30%. We consider our metric to be the ratio

$$\frac{\#\text{ of cuts}}{\#\text{ of expected cuts}} \tag{2.13}$$

where the # of expected cuts is 20 because our tested time series have total length $N = 4000$ with segments of $m = 200$ and the # of cuts is the number of cuts the segmentation algorithm performs on the tested Cauchy time series with scale parameters γ_1, γ_2. The ratio obtained from Eq. (2.13) indicates the performance of the algorithm for each set of parameters γ_1, γ_2, taking the arithmetic mean of the collection of ratios obtained from both tests we get that the KS test performs, in average, with 83.6% while the AD has an average performance of 88.3%. Hence we can state that the segmentation algorithm with the AD modification performs better than with the KS test. In Table 2.5 are summarised the counts and percentage of the results obtained in Fig. 2.25.

Analysing Table 2.5 we conclude that the AD implementation leads to more oversegmentation and less undersegmentation comparing with the KS variant of the segmentation algorithm. Also, the KS test seems to cut correctly the time series more frequently than

TABLE 2.5: Summary of results for the comparison of the performnce of the KS segmentation algorithm between KS and AD based tests. The table shows the number of occurences and the matching percentage that correspond to oversegmentation with more than 20 and more than 25 cuts, indicating extreme oversegmentation, undersegmentation with under 20 and under 10 cuts, indicating extreme undersegmentation, and the number of correct segmentation with exactly 20 cuts. The two last rows show extreme situations of over, > 25, and undersegmentation, < 10. The table translates the main features observed in Fig. 2.25, overall that for the segmentation algorithm based on the KS test we obtained more undersegmentation and more correct cuts while with the AD implementation the algorithm carries out more oversegmentation.

	KS		**AD**	
	#	%	#	%
$>$**20**	228	57.0	252	63.0
$<$**20**	132	33.0	121	30.3
$=$**20**	40	10.0	27	6.8
$>$**25**	22	5.5	56	14
$<$**10**	108	27.0	102	25.5

the AD test as we can see in the third row of Table 2.5, 10% for KS against 6.8% for AD.

Chapter 3

Discussion and Conclusion

3.1 Discussion

Starting from an automated segmentation algorithm based on the Kolmogorov-Smirnov (KS) distance for Gaussian distributed random time series [3], we have attempted to characterise and improve the segmentation performance for heavy tailed time series. In a primary phase we made a complete characterisation of the of standard Kolmogorov-Smirnov (KS) test, in Section 2.1 where we explored the KS probability function in Fig. 2.1 for large sample sizes and the significance criterion for the classic KS test in Fig. 2.2 showing that, for large N, the curves tend to 1.22, 1.36 and 1.63, for the respective confidence levels. To complete the primary analysis of the classic KS test we look at the ability of the test to determine whether two samples are drawn from the same distribution or not. Given two time series of length N we plotted the KS probability function, Q_{KS}, and the distance, D against N illustrated in Fig. 2.4 from where we conclude that the two sample KS test can safely reject the null hypothesis that the two distributions are the same if the significance is sufficiently high and N is large enough, Fig. 2.4 (B). However, this does not imply that the test is capable of deciding the opposite question, whether the two distributions are the same, Fig. 2.4 (A). After the characterisation of the classic KS test we introduced in Section 2.3 the KS based segmentation algorithm described in Section 1.7, where we performed numerical tests to evaluate the accuracy of the algorithm in detecting differences within an artificial time series running the KS segmentation algorithm once on different samples with known parameters, specified in Table 2.1, using a suitable significance criterion given by Eq. 2.3 [3], which is more restrictive than the classic one, Eq. 1.22, this new criterion is explored in Section 2.4 . The results are shown in Fig. 2.5 and indicate that the KS segmentation algorithm

is accurate in detecting differences within a time series composed of Gaussian random numbers.

Since the Q_{KS} criterion, given by Eq. 1.22 and shown in Fig. 2.2 is derived under the assumption that we have independent time series and this is not the case when we apply segmentation within one time series, in fact, all segments will be interdependent and for this reason we need a more restrictive criterion, the use of the classic KS test would lead to oversegmentation. This criterion was derived in the work of [3], given by Eq. 2.3 and was computed and studied in Section 2.4.

We then looked at answering the question "How well does the KS test work in detecting meaningful differences within a sample composed of fragments with different distributions?" and concluded that it works surprisingly well for the Gauss, log-normal and Cauchy distributions without any modification to the test. Also we introduced a minimum length parameter l_0, following the approach presented in [3], to be evaluated before the algorithm makes a cut, without this requirement, i.e. $l_0 = 0$ there is considerably more over segmentation that occurs. By setting a larger value of l_0, smaller segments are discarded and the number of false cuts is reduced as can be seen in Figs. 2.13, 2.16 and 2.17. In [3] the KS segmentation algorithm was only applied to Gaussian distributions When testing the performance of the test for Gauss, log-normal and Cauchy distributions we were motivated to seek for improvements for tailed distributions which was done by means of the Anderson-Darling (AD) modification. Since the AD test gives more weight to the tails than does the KS test we expected that the AD modification to the KS test would improve significantly the performance of the segmentation algorithm for heavy tailed distributions, namely the Cauchy distribution. The comparison between KS and AD tests are shown in Fig. 2.25 where we observe a slight but not too significant increase in performance, undersegmentation occurs less frequently for the AD implementation, 30.3%, than for the KS, 33%, showing an improvement of 2.7% in undersegmentation. On the other hand, the AD test shows increased oversegmentation, 63% compared to the the 57% observed for the KS test both tested on Cauchy distributed time series in Fig. 2.25.

3.2 Conclusion

In this thesis we presented a segmentation algorithm which aims at coping with non stationary time series based on the work of [3]. The algorithm is based on the KS test which shows itself accurate at detecting differences of CDF's when dealing with distributions that do not have relevant information in the tails. This motivated us to look for an improvement that would work well for heavy tailed distributions where we applied a

modification to the KS test introducing a weighting term to the KS distance which is called the Anderson-Darling test, we have noticed a slight but not significant improvement in the performance of the algorithm. Further work should consist in improving the weight term and look to improve sensitivity near the tails. Here the tasks were focused in evaluating and testing the algorithm extensively to artificial time series and as further work the present algorithm should be applied to empirical measurements from complex physical, geophysical or socio-economical systems, where heavy tailed distributions often play a crucial role. A similar study was done in literature in [3], [5] where the authors tested the KS segmentation algorithm on a fragment of a heart rate time series, a series of wind velocities for one month of measurements at a 30 s acquisition interval and it is shown that is works well in detecting differences within a time series with mixed statistics.

It will be interesting to see how the modified algorithm presented here can help distinguish different parameter regimes in real-time complex series, for example in financial market data, where typically oscillations occur between different market states or sentiments, accompanied by changes in return distributions, correlation structure or Hurst exponents, among others. Possibly in combination with other statistical tools sensitive for changes in the latter quantities, an automated segmentation routine can be a helpful, fast and easily programmable aid in decision making.

Appendix A

Iteration Pseudo Code

```
while(length_sites_to_visit > 0)
{

     sites_to_visit = [[0, N]]
     sites_visited = []

     first iteration: add ptr1, 0 < ptr1 < N

     sites_to_visit = [[0, ptr1], [ptr1, N]]
     sites_visited = [[0, N]]

     second iteration: add ptr2, 0 < ptr2 < ptr1

     sites_to_visit = [[0, ptr2], [ptr2, ptr1]]
     sites_visited = [[0, N], [0, ptr1], [ptr1, N]]

     third iteration: no cut performed

     sites_to_visit = []
     sites_visited = [[0, N], [0, ptr1], [ptr1, N], [0, ptr2], [ptr2, ptr1]]

}
```

Bibliography

[1] Frank Raischel, Teresa Scholz, Vitor V. Lopes, and Pedro G. Lind. Uncovering wind turbine properties through two-dimensional stochastic modeling of wind dynamics. *Physical Review E*, 88(4):042146, 2013. doi: 10.1103/PhysRevE.88.042146.

[2] Paulo Rocha, Frank Raischel, João P da Cruz, and Pedro G Lind. Stochastic evolution of stock market volume-price distributions. *arXiv preprint arXiv:1404.1730*, 2014.

[3] S. Camargo, S. M. Duarte Queirós, and C. Anteneodo. Nonparametric segmentation of nonstationary time series. *Physical Review E*, 84(4):046702, 2011. doi: 10.1103/PhysRevE.84.046702.

[4] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, second edition, 1992.

[5] Pedro Bernaola-Galván, Plamen Ivanov, Luís Nunes Amaral, and H. Stanley. Scale invariance in the nonstationarity of human heart rate. *Physical Review Letters*, 87 (16):168105, 2001. doi: 10.1103/PhysRevLett.87.168105.

[6] Hisashi Kobayashi, Brian L Mark, and William Turin. *Probability, Random Processes, and Statistical Analysis: Applications to Communications, Signal Processing, Queueing Theory and Mathematical Finance*. Cambridge University Press, 1 edition, 2011.

[7] Roy D Yates and David J Goodman. *Probability and Stochastic Processes: A Friendly Introduction for Electrical and Computer Engineers*, volume 2. John Wiley & Sons, 2005.

[8] Catherine Forbes, Merran Evans, Nicholas Hastings, and Brian Peacock. *Statistical distributions*. John Wiley & Sons, 4 edition, 2011.

[9] Mark EJ Newman. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[10] Hwei P Hsu. Theory and problems of probability, random variables, and random processes. *Schaum's Outline Series*, 1997.

[11] Richard Phillips Feynman, Robert B Leighton, and Matthew Sands. *Feynman lectures on physics. vol. 1: Mainly mechanics, radiation and heat.* Addison-Wesley, 1963.

[12] Myles Hollander, Douglas A Wolfe, and Eric Chicken. *Nonparametric Statistical Methods.* John Wiley & Sons, third edition, 2013.

[13] Wikipedia, The Free Encyclopedia. Kolmogorov-smirnov test. `https://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test`, 2014. [Online; accessed 18-December-2014].

[14] Peter Sprent and Nigel C. Smeeton. *Applied Nonparametric Statistical Methods.* Chapman and Hall/CRC, 4 edition, 2007.

[15] Andrey Nikolaevich Kolmogorov and BV Gnedenko. Limit distributions for sums of independent random variables. *Addison-Wesley, Cambridge, Mass*, 1954.

[16] Benoit Mandelbrot. The pareto-levy law and the distribution of income. *International Economic Review*, 1(2):79–106, 1960.

[17] Svetlozar Rachev and Stefan Mittnik. *Stable Paretian models in finance.* John Wiley & Sons, 1 edition, 2000.

[18] Wikipedia, The Free Encyclopedia. Heavy-tailed distribution. `https://en.wikipedia.org/wiki/Heavy-tailed_distribution`, 2015. [Online; accessed 25-June-2015].

[19] Theodore W Anderson and Donald A Darling. Asymptotic theory of certain "goodness of fit" criteria based on stochastic processes. *The annals of mathematical statistics*, pages 193–212, 1952.

[20] David M Mason and John H Schuenemeyer. A modified kolmogorov-smirnov test sensitive to tail alternatives. *The Annals of Statistics*, pages 933–946, 1983.

[21] Rémy Chicheportiche and Jean-Philippe Bouchaud. Weighted kolmogorov-smirnov test: Accounting for the tails. *Physical Review E*, 86(4):041115, 2012.

[22] The Scipy community. Two-sample kolmogorov-smirnov test. `http://docs.scipy.org/doc/scipy-0.14.0/reference/generated/scipy.stats.ks_2samp.htm` 2014. [Online; accessed 31-March-2016].

[23] Wikipedia, The Free Encyclopedia. Cauchy distribution. `https://en.wikipedia.org/wiki/Cauchy_distribution`, 2016. [Online; accessed 16-February-2016].

[24] B Everitt. *The Cambridge dictionary of statistics*. Cambridge University Press, 3 edition, 2002.